Three real-world datasets and neural computational models for classification tasks in patent landscaping

Subhash Pujari, Jannik Strötgen, Mark Giereth, Michael Gertz, Annemarie Friedrich

Angaben zur Veröffentlichung / Publication details:

Pujari, Subhash, Jannik Strötgen, Mark Giereth, Michael Gertz, and Annemarie Friedrich. 2022. "Three real-world datasets and neural computational models for classification tasks in patent landscaping." In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 7-11 December 2022, Abu Dhabi, United Arab Emirates*, edited by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, 11498–513. Stroudsburg, PA: Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.791.

Three Real-World Datasets and Neural Computational Models for Classification Tasks in Patent Landscaping

Subhash Chandra Pujari[1,3] **Jannik Strötgen**[1] **Mark Giereth**[2]
Michael Gertz[3] **Annemarie Friedrich**[1]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Robert Bosch GmbH, Stuttgart, Germany
[3]Institute of Computer Science, Heidelberg University, Heidelberg, Germany
firstname.lastname@de.bosch.com
gertz@informatik.uni-heidelberg.de

Abstract

Patent Landscaping, one of the central tasks of intellectual property management, includes selecting and grouping patents according to user-defined technical or application-oriented criteria. While recent transformer-based models have been shown to be effective for classifying patents into taxonomies such as CPC or IPC, there is yet little research on how to support real-world Patent Landscape Studies (PLSs) using natural language processing methods. With this paper, we release three labeled datasets for PLS-oriented classification tasks covering two diverse domains. We provide a qualitative analysis and report detailed corpus statistics.

Most research on neural models for patents has been restricted to leveraging titles and abstracts. We compare strong neural and non-neural baselines, proposing a novel model that takes into account textual information from the patents' full texts as well as embeddings created based on the patents' CPC labels. We find that for PLS-oriented classification tasks, going beyond title and abstract is crucial, CPC labels are an effective source of information, and combining all features yields the best results.

1 Introduction

A patent is a public document granting the exclusive rights to an invention, e.g., a product or a process that provides a new technical solution to a problem. When entering new markets or developing new products, it is of utmost importance for organizations such as companies to be aware of the **patent landscape**, i.e., the existing patents with regard to their business endeavor in order to ensure their freedom to operate. Experiments conducted with expert patent examiners in a context of a feasibility study on prior art search showed that while fully automating the process is infeasible, Natural Language Processing (NLP) methods can help to significantly reduce time and cost by assisting patent examiners (Setchi and Spasic, 2020).

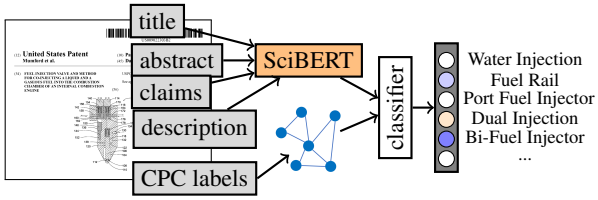


Figure 1: A **Patent Landscape Study** commonly involves the classification of patents into a set of business- or application-oriented **target classes**.

The task of obtaining an overview of the relevant intellectual property with the aim of supporting strategic decisions is also called performing a **Patent Landscape Study** (**PLS**). A PLS consists of three steps: search, classification, and analysis. The first step identifies relevant documents that are then grouped into a set of user-defined categories. Finally, the labeled dataset is used to derive essential insights. Most efforts towards automating the PLS process (Abood and Feltenberger, 2018; Choi et al., 2022) focus on the first step of the process. In contrast, in this work, we address document classification, i.e., the second step of the PLS process.

Most prior work on patent classification engages in the task of assigning labels from the International Patent Classification (IPC) taxonomy and its extension, the Co-operative Patent Classification (CPC) taxonomy (Smith, 2002; Grawe et al., 2017; Li et al., 2018; Lee and Hsiang, 2020; Pujari et al., 2021). This hierarchical multi-label classification task is challenging due to the large number of CPC codes, but lots of training data exists as each patent gets assigned CPC labels upon submission to the patent office. CPC classification provides a good testbed for developing representations of patents.

However, PLSs are performed on a set of patent applications or granted patents with the aim of categorizing the documents according to a set of application- or business-oriented criteria, which may correspond to CPC categories only to a lim-

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11498 - 11513

ited extent. To clearly differentiate from prior work on CPC classification, we call this PLS-oriented task the *target classification task* (see Figure 1). In our setting, documents are already labeled with CPC labels, and thus these labels can be leveraged as one source of information. Despite the expected major impact on the speed and accuracy of PLSs, NLP research on such PLS-oriented target classification tasks has been hindered by the unavailability of public datasets with exemplary tasks.

The first important contribution of this work is to release three real-world datasets from two diverse domains, providing a testbed for developing PLS-oriented classifiers.[1] We release a large manually curated patent corpus, which has been annotated with target labels related to injection valves during a time period of 20 years by domain experts of an industrial collaborator. In addition, we enrich two smaller document collections from WIPO, created during real-world PLSs on HIV drugs, and define benchmark tasks on them. We provide a detailed analysis and corpus statistics, highlighting the difficult nature of the target classification tasks due to class imbalance and multi-label scenarios.

Our second main contribution is a computational study to tackle the target classification, comparing recently proposed neural and non-neural models that have been shown to be effective for CPC classification. Building on the work of Choi et al. (2022), and inspired by Pujari et al. (2022), we experiment with combinations of content- and label-based feature vectors. We generate SciBERT-based (Beltagy et al., 2019) embeddings for the patents' title and abstract, claims, and description. To represent the semantics of CPC labels, we compare different approaches to generate embeddings based on a label co-occurrence graph and the label descriptions' texts. We find that using all textual fields as well as the CPC embeddings results in a robust method that works consistently well across our three PLS datasets, outperforming all baselines.

In sum, our contributions are as follows:

- We **define the novel task** of PLS-oriented *target classification* and provide **three real-world datasets** as benchmarks.
- Our in-depth **corpus study** details the nature of the datasets as well as the target tasks.
- In our **computational study**, we propose a **robust architecture** that works well across all three datasets.
- We show that across datasets, good accuracy (micro-F1) can be reached by only annotating about 200 samples, but that further research is needed to boost performance in the long tail.

[1] https://github.com/boschresearch/pls_benchmark_emnlp2022

2 Related Work

We group our review of related work into patent classification, automated patent landscaping, and metadata-based patent document representations.

Patent Datasets. For CPC classification, various datasets are available (Pujari et al., 2021, 2022; Li et al., 2018). Similar to our target classification task, Richter and MacFarlane (2005) study classification for a patent alert system for the biochemical domain, but the dataset was not open-sourced. Sharma et al. (2019) provide a patent dataset with human-written abstractive summaries. In the context of prior-art search, Risch et al. (2020) release a dataset mapping claims to prior-art passages.

Patent Classification. Early patent classification systems (Fall et al., 2003; Guyot et al., 2010; Wu et al., 2010; Verberne and D'hondt, 2011) use a TF-IDF feature vector exploiting the full document text. CNNs (Li et al., 2018; Niu and Cai, 2019), RNNs (GRU (Risch and Krestel, 2019), and LSTMs (Grawe et al., 2017)) have also been used to represent patent text. Recently, pre-trained transformer-based models have been shown to be effective for patent classification (Lee and Hsiang, 2020; Pujari et al., 2021; Althammer et al., 2021). Transformer-based models are constrained to a maximum length of 512 input tokens. Increasing the maximum sequence length to 4096 tokens, Zaheer et al. (2020) propose Big Bird, a long-text transformer, and apply it to CPC classification using the concatenated text of title, abstract, and claims as input. As these approaches are inefficient (Park et al., 2022) and to date have shown only limited improvements over RoBERTa (Liu et al., 2019) for patent classification, we leave research on long-transformer methods to future work.

Automating Patent Landscaping. We are aware of several works aiming at automating the first step of the PLS process. Abood and Feltenberger (2018) first identify relevant patents by expanding a seed list, performing forward and backward traversal on the patent citation graph, also relying on the relevant CPC labels. They then train a classifier using one-hot embeddings of references and CPC codes, as well as an embedding of the patent abstract using word2vec and an LSTM to

predict whether a patent is relevant to a PLS or not. Similarly, for a PLS in the artificial intelligence domain, Giczy et al. (2022), propose a classifier concatenating the LSTM output for abstract and claims to the citation embedding. Choi et al. (2022) employ a model architecture more similar to ours, using a transformer to embed the abstract and the graph neural network diff2vec (Rozemberczki and Sarkar, 2018) to embed CPC labels. In contrast to these works, we target the second step of the PLS process, categorizing a set of retrieved documents into business- or application-oriented categories.

Embedding Metadata for Patent Classification. In the contexts of classification (Richter and MacFarlane, 2005; Benites et al., 2018) and clustering (Vlase et al., 2012), non-neural count- and TF-IDF-based feature vectors reflecting IPC, inventor, and assignee information have been proposed. For CPC classification, Niu and Cai (2019) leverage the BM25-similarity between the document text and the CPC label descriptions. Fang et al. (2021) compute embeddings over graphs constructed from word co-occurrence, inventor, and assignee information, and combine them using attention-based sums. In contrast, we decide to restrict our study to content-based features, as using inventor and assignee information might introduce biases that contradict with the goal of a PLS.

3 Patent Landscaping: Task and Datasets

In this section, we propose a *target* classification task to support PLS, and introduce three new data sets from diverse domains (mechanical systems and biochemistry). We have curated one dataset from in-house annotations of patents in the domain of injection valves and compiled two datasets from freely available WIPO patent landscape studies. All datasets are publicly available for future benchmarking in a convenient format. Dataset statistics are provided in Table 1. Label distributions are shown in Figure 2.

3.1 Target Classification Task

Given the training data $\{\langle d^{(i)}, \boldsymbol{C}_d^{(i)}, \boldsymbol{L}_{\boldsymbol{d}}^{(\boldsymbol{i})}\rangle\}_{i=1}^{n}$, the target classification task estimates a function mapping a document $d^{(i)}$ that comes with a set of CPC labels $\boldsymbol{C}_d^{(i)}$ to a target label set $\boldsymbol{L}_d^{(i)}$. A patent document d consists of the textual fields title (t), abstract (a), claims (cl), and description ($desc$). The CPC labels $\boldsymbol{C}_d$ are taken from the predefined CPC taxonomy, which also provides a textual description for each label.[2] The target label set $\boldsymbol{L}_d$ contains the user-defined application- or business-oriented categories relevant to the current PLS.

Dataset	# Instances	# Unique Labels	Avg. # Labels Per Instance
InjVal	9,465	16	1.01
Rito	781	7	1.35
Atz	640	8	2.14

Table 1: **Target label statistics** of PLS datasets.

3.2 Injection Valves Dataset

In the **InjVal** dataset, patent families are labeled with categories describing types of injection valves and related technologies. The dataset has been labeled by an in-house domain expert, a patent attorney and expert in injection valves with over 30 years experience in the related IP management, who performed the classification task on a weekly basis for the past 25 years. Each week, a candidate set of patents is generated by an alert system that filters the new incoming patents using a CPC-based search query. The domain expert identifies relevant patents in the candidate set, and categorizes them into a technical target category. Since most patents belong to mechanical systems, the domain expert often made use of the patents' figures when making relevance judgments.

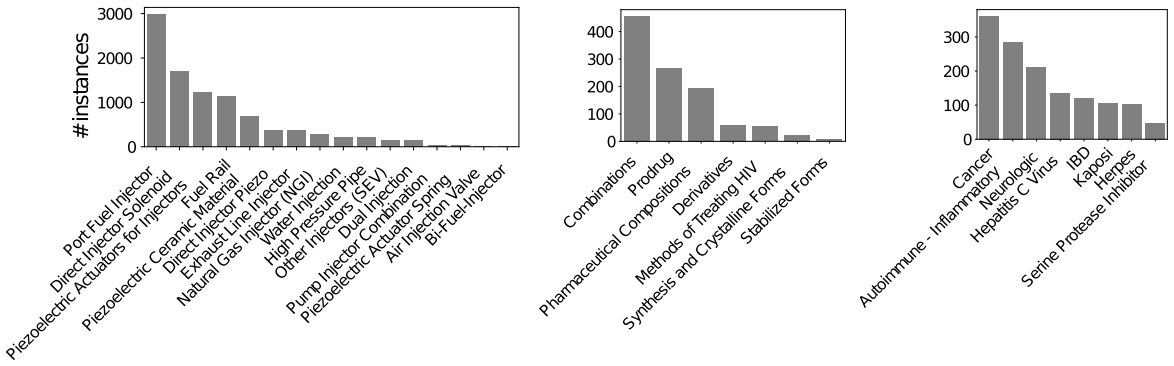
The 9,465 patent families are labeled with 16 different target labels indicating the injector components or injection types. The majority of patents are from the Japanese and German Patent Offices, followed by US patents (see appendix A.2, Figure 6a). We add the corresponding English machine-translated text for each field.[3] The dataset covers a broad domain (5,068 CPC labels) and hence corresponds to a higher-level PLS. The average number of labels per instance is close to 1, resulting in a single-label classification task.

3.3 Ritonavir and Atazanavir Datasets

We derive two labeled datasets from two publicly available PLSs by the World Intellectual Property Organization (WIPO) on Ritonavir (**Rito**) and Atazanavir (**Atz**), two drugs developed for the treat-

[2] We use IPC/CPC labels, but only refer to CPC labels for readability in the following.

[3] We thank PatBase (/www.patbase.com), RWS (www.rws.com) and MineSoft (https://minesoft.com/) for agreeing to the publication of the translated texts.



(a) Injection-Valve (b) Ritonavir (c) Atazanavir

Figure 2: **Label distributions** in the PLS-oriented target classification task datasets.

ment of HIV infections and AIDS.[4] The motivation of the studies, both conducted in 2011, was to track the development of the drug manufacturing process as well as the drugs' compositions and usage since the filing of the first invention. In contrast to InjVal, these two datasets contain patent families within a narrow scope about a single invention.

The PLSs have been conducted within WIPOs Development Agenda project "Developing Tools for Access to Patent Information," which aims to research and describe the patterns of patenting and innovation activity related to specific technologies.[5] For each report, WIPO collaborated with institutional partners working in the respective field and having an interest in the specific topic. The search methodology was documented carefully. The report on Atazanavir was conducted by the Thomson Reuters IP Solutions and IP Consulting Group in cooperation with the Medicines Patent Pool (MPP). The report on Ritonavir has been compiled by Landon IP.

The WIPO studies have been conducted in an iterative manner. First, a keyword-based search yielded a list of relevant documents, which was then filtered using relevant CPC labels. With a forward-backward citation search, some additional patents have been identified and added to the dataset. Each study provides a spreadsheet-like overview with meta-information about the search and patents, for instance, labels have been assigned to the patents which correspond to classification target labels performed during a PLS. The labels have been carefully assigned by WIPO professionals during search (for Rito) and post-hoc supported by text mining software[6] (in the case of Atz). In contrast to the gold-standard InjVal and Rito datasets, the labels in Atz should thus rather be considered as silver standard. By analysing the descriptions within the reports, we select subsets of these labels as PLS target labels for our experimental studies.

The Atz data as provided by WIPO contains the title and an abstract by Derwent[7] together with the first claim. The Rito data only lists title, abstract, and claims. As part of our contribution, we derive a structured full-text dataset from the information provided by WIPO by adding additional information from PatBase. In an easy-to-use format, we provide title, abstract, (all) claims, the description text, CPC labels, the patent number, the family number, and the publication date.

The **Rito** dataset consists of 781 patent families, labeled with seven distinct target labels. These correspond to broad categories that have been assigned during search by carefully choosing queries based on keywords such as disease names or chemical compositions in combination with CPC classes. The categories include *Methods of Treating HIV*, and *Combination* and *Prodrug*, which relate to the methods of administering the drug. The remaining four categories (*Pharmaceutical Composition*, *Derivatives, Synthesis and Crystalline Forms*, and *Stabilized Forms*) define the form, composition, and derivatives of Ritonavir.

The **Atz** dataset consists of 640 patent families and eight target labels, which are the names of the type of disease whose treatment is described in the patent. While the primary indication of

[4] https://www.wipo.int/publications/en/details.jsp?id=230 https://www.wipo.int/publications/en/details.jsp?id=265

[5] https://www.wipo.int/edocs/mdocs/mdocs/en/cdip_4/cdip_4_6.pdf, DA_19_30_31_01

[6] thevantagepoint.com/6-products/thomson-data-analyzer.html

[7] clarivate.com/derwent/solutions/derwent-world-patent-index-dwpi

Field	InjVal	Rito	Atz
abstract	104 ± 43	60 ± 35	56 ± 34
claims	358 ± 346	1215 ± 1051	1231 ± 948
description	2121 ± 1171	11579 ± 8800	16401 ± 10245

Table 2: **Token Counts**: Mean and standard deviation by dataset and textual field.

Dataset	Documents	Unique Labels	Labels Per Instance
InjVal	9465	5068	6.42
Rito	781	3543	26.87
Atz	640	3171	31.18

Table 3: **CPC/IPC Statistics** of PLS datasets.

Atazanavir is HIV, medical professionals have administered it for other indications, and we select the subset of patents that describe a non-HIV indication, defining the target task as identifying the corresponding (non-HIV) disease.[8] Among the target labels, *Cancer* is the most frequent one, followed by *Autoimmune-Inflammatory*.

While both Atz and Rito focus on HIV-related drugs, the two PLS tasks are qualitatively different: Rito divides patents by technology, Atz divides patents by application. As shown in Table 1, the average label per instance is larger than 1, i.e., Rito and Atz constitute multi-label classification tasks.

3.4 Dataset Analysis and Corpus Statistics

The characteristics of a dataset affect the performance of classification models. To allow for a better interpretation of our experimental results, we first perform a statistical analysis of the datasets.

Token Counts / Text Lengths. We tokenize the texts of all patent fields using the NLTK whitespace tokenizer and report average token counts in Table 2. The abstracts in InjVal are longer compared to those of Rito and Atz, which have longer claims and description sections. Also, we see a high variation in the token count, in particular within the description section for Rito and Atz.

Publication Date. The publication date of a patent family is the earliest publication date among its family members. The InjVal dataset covers patent families with a broader time horizon of around 100 years, while Atz and Rito contain patents within a shorter period of 16 and 22 years, respectively (see appendix A.1, Figure 5).

Patent Office / Original Language. For Rito and Atz, most patents are from USPTO or have a worldwide filing through WIPO (see appendix A.2, Figure 6). Thus, respective patents are written in English. For InjVal, the majority (68%) of patents consist of machine-translated text.

Duplicate Abstracts. As patent abstracts are not legally binding, companies often re-use the same abstracts, sometimes to consciously conceal information. In our datasets, there are 48, 109, and 90 patents in InjVal, Atz and Rito, respectively, that do not have unique abstracts. Some abstracts in InjVal and Rito occur up to 20 times (for more details, see appendix A.3, Figure 7). This illustrates why methods based only on abstracts are suboptimal.

CPC Labels. Table 3 shows the CPC statistics. The patents within the WIPO datasets have a higher number of labels compared to the InjVal dataset. The InjVal dataset contains only one patent with a CPC count of more than 50, whereas Rito and Atz contain 13 and 18 such patents, respectively. Also, the numbers of unique CPC labels within the WIPO datasets are comparatively higher given the relatively smaller sizes of the datasets. We hypothesize that the effectiveness of using CPC labels as features depends on the correspondence between CPC and the target labels. In our analysis (Appendix B) comparing the Pointwise Mutual Information (Church and Hanks, 1990) between CPC and target labels, we observe a higher similarity between CPC and target labels in InjVal than in Rito and Atz. Compared to Atz, target labels in Rito have a higher correspondence to CPC labels.

4 Computational Models

In this section, we describe our computational models for predicting target categories for patents based on their text and CPC labels. We first introduce the representations of the patent text (Section 4.1), as well as the generation of embeddings for CPC labels (Section 4.2), and then describe the classifier used on top of them (Section 4.3).

4.1 Neural Patent Text Representations

For a given textual field, we generate a sequence of word-piece tokens, truncate it to a maximum sequence length of 510, and pass it through SciBERT (Beltagy et al., 2019), a BERT-style text encoder (Devlin et al., 2019) pre-trained on scientific text.

[8] Despite this definition of target labels in Atz, the correct classification of patents cannot be easily achieved with keyword-based approaches and our more sophisticated, robust approach (cf. Section 4) highly outperforms such simple baselines. See Appendix D for further details.

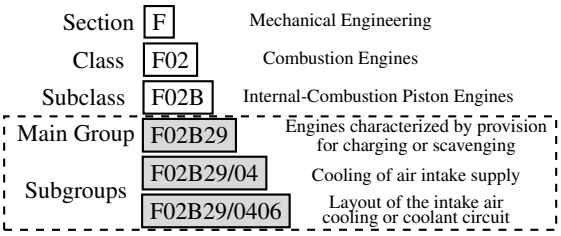


Figure 3: CPC scheme with label descriptions.

Althammer et al. (2021) found that SciBERT-based models outperform BERT on a CPC classification task. We use the last hidden state of [CLS] token as the text's embedding, denoted by $e(.)$. We fine-tune a model on the CPC classification task described by Pujari et al. (2021), and then use the resulting fine-tuned SciBERT model to compute embeddings in all of our experiments. We compute three text embeddings: $e(t+a)$ using the concatenated text of title and abstract; $e(cl)$ using the claims' text; and $e(desc)$ using the text of the description. When using them jointly, we use vector summation ($\oplus$) following Pujari et al. (2022).

4.2 CPC Label Embeddings

We experiment with four different ways of embedding knowledge about the CPC labels associated with a patent document. The simplest embedding consists of a **multi-hot encoded vector** ($cpc_{multihot}$) with each dimension indicating the presence of one CPC label.

Each CPC label comes with a textual description. For example (see Figure 3), the description of the label F02B29 is "Engines characterised by provision for charging or scavenging." For each CPC label, the full description is generated by traversing the path from the respective main-group node, concatenating the label descriptions at each hop. For example, the label description for F02B29/0406 is the concatenation of the descriptions of F02B29, F02B29/04, and F02B29/0406. We then compute a **SciBERT-embedding for the description text** as described in Section 4.1. The document-level embedding cpc_{text} is the mean over the CPC label embeddings of all CPC classes assigned to the patent. Further, since label descriptions contain important domain-specific keywords, we compute a 140k-dimensional **TF-IDF feature vector** $cpc_{tf.idf}$ **for the concatenated label-description texts** of the CPC labels assigned to a document using a TF-IDF model computed over all the label description within the CPC taxonomy.

In addition, we compute **graph embeddings for CPC labels**. For this, we construct a graph with all CPC labels that occur in our datasets as nodes. Pairs of nodes are connected if the corresponding CPC labels co-occur in a document. Edge weights correspond to the co-occurrence count of the two CPC labels. To generate the label embeddings, we use the node2vec algorithm proposed by Grover and Leskovec (2016), employing the StellarGraph (Data61, 2018) implementation. The algorithm performs multiple random walks, generating random biased node sequences (influenced by the edge weights) that are fed into a word2vec model (Mikolov et al., 2013), which then computes the node embeddings. The document-level cpc_{graph} embedding is the mean of the embeddings of the document's CPC labels.

4.3 Classification Model

Our classification model architecture is similar to the Transformer-based Multi-task Model (TMM, Pujari et al., 2021). As input to the model, the CPC-label and patent-text based embeddings are either used in isolation or combined using vector concatenation (;). The TMM model employs one classification head for each label. Each head consists of three dense layers. The last dense layer has a binary softmax output, predicting whether or not a label applies.

5 Experiments

In Section 4, we have introduced several content and label embeddings for PLS target classification tasks. Our experiments described in this section aim to identify a patent document representation that works robustly across PLSs. We analyze the performance of different embeddings when considered individually or in combination (Section 5.3), and compare it to strong baselines (Section 5.4). Finally, as an important analysis in the context of PLSs, in Section 5.5, we address the question of how many labeled training examples are necessary for training a PLS target task classifier. Details about hyperparameters for our proposed approach are provided in appendix C.

5.1 Baselines

We compare our approach against state-of-the-art neural and non-neural models.

TMM with $e(t+a)$. As a neural baseline, we use the setup as proposed by Pujari et al. (2021)

Model	InjVal		Rito		Atz	
	macro-F1	**micro-F1**	**macro-F1**	**micro-F1**	**macro-F1**	**micro-F1**
Benites et al. (2018): SVM	61.4±2.1	74.1±4.3	51.1±6.8	58.2±2.5	65.4±1.8	71.9±2.5
Pujari et al. (2021): TMM + $e(t+a)$	65.2±2.1	79.2±1.3	44.3±4.0	66.0±3.0	62.1±2.2	70.6±1.1
TMM + $e(t+a) \oplus e(cl)$	66.1±2.0	82.0±0.8	39.3±5.1	64.5±1.8	64.7±1.8	71.3±2.1
TMM + $e(t+a) \oplus e(cl) \oplus e(desc)$	66.2±4.9	82.2±1.7	49.1±6.9	66.3±1.9	62.6±4.0	71.2±2.6
TMM + $cpc_{multihot}$	49.8±3.7	77.8±0.9	17.3±2.5	42.0±8.0	23.8±5.3	39.9±5.2
TMM + cpc_{text}	54.7±2.8	73.8±0.5	28.1±1.9	60.5±2.5	37.0±3.5	47.9±1.5
TMM + cpc_{graph}	58.6±1.7	76.5±0.7	35.2±5.5	62.9±1.9	44.2±3.2	50.8±3.8
TMM + $cpc_{tf.idf}$	60.4±1.9	75.9±0.8	39.2±6.0	60.5±3.9	44.4±2.5	52.8±2.2
SVM + $cpc_{tf.idf}$	63.0±1.2	76.7±1.4	45.2±5.4	61.4±2.1	50.1±1.9	58.6±1.2
TMM + $e(t+a) \oplus e(cl) \oplus e(desc)$; $cpc_{tf.idf}$	66.6±0.5	83.9±0.4	46.4±5.7	65.1±2.6	63.4±2.8	71.1±1.1
TMM + $e(t+a) \oplus e(cl) \oplus e(desc)$; cpc_{graph}	**67.7**±2.5	**84.3**±0.5	**53.9**±6.6	**67.7**±3.2	**66.2**±1.8	**73.2**±2.1

Table 4: Comparison of text-based and CPC-based embeddings. Benites et al. (2018) uses TF-IDF-based vectors for title, abstract, description, and claims.

for multi-label classification with a Transformer-based Multi-task Model (TMM). Document representations correspond to the $e(t+a)$ method using SciBERT.

SVM. Conceptually simpler *term frequency–inverse document frequency* (TF-IDF) vectors are still often used in text classification tasks (Malmasi et al., 2016; Sulea et al., 2017; Benites et al., 2018). They often show surprisingly strong performance despite their simplicity, likely because they can easily incorporate information from long documents. For instance, in the context of the ALTA 2018 shared task on multi-label IPC classification, Benites et al. (2018) achieved competitive results with a support vector machine (SVM, Cortes and Vapnik, 1995) ensemble-based approach. The 140k-dimensional feature vector for the complete document text, i.e., $t+a+cl+desc$, comprises TF-IDF values for 70k character n-grams (3- to 6-chars) and 70k word n-grams (1- to 2-grams).

5.2 Dataset Splits

We divide each dataset into two parts. The heldout test set is a sample that a model has never seen during training and contains 15% of the total instances. The remaining 85% of the data are used for 5-fold cross-validation (CV), which we use to tune the models. For each cross-validation fold, we use three folds as our training set, one fold for tuning, and one as dev set. Finally, each of the five models is evaluated on the test set. We report the mean and standard deviation values across these five evaluations.[9]

5.3 Comparison of Patent Embeddings

The upper part of Table 4 reports scores for using various combinations of the **patent text embeddings**. For InjVal, we see consistent improvements when adding $e(cl)$ and $e(desc)$. Adding $e(cl)$ leads to mixed results on Rito and Atz, however, with the exception of macro-F1 for Atz, using all three text embeddings at once performs generally well.

The middle part of Table 4 shows results for using various **CPC-label based embeddings**. Among these, SVM+cpc_{tfidf} achieves the best macro-F1 scores and the highest micro-F1 scores for InjVal and Rito. The patent-text based embeddings outperform the best model using only CPC information (SVM+cpc_{tfidf}). Among the more sophisticated CPC label feature vectors, TF-IDF with the concatenated label descriptions ($cpc_{tf.idf}$) performs best across datasets in terms of macro-F1. Comparing the neural label embeddings across datasets, we observe that the graph-based embeddings (cpc_{graph}) consistently outperform the description-based embeddings (cpc_{text}).

Finally, as a sanity check to demonstrate that there is no one-to-one mapping between target labels and CPC labels in our proposed datasets, we evaluate the performance with multi-hot encoded vector $cpc_{multihot}$ as the only feature. As expected due to the analysis in Section 3.4, performance for InjVal is higher than for the WIPO datasets.

When **combining CPC embeddings with the patent-text embeddings** in TMM, cpc_{graph} outperforms cpc_{tfidf}. While cpc_{graph} is directly trained as a dense embedding, combining cpc_{tfidf} with the TMM model is not straightforward due to its high

[9]The relatively high standard deviations we report result from using slightly *different* training sets in each of the five folds. We use this setting because it leads to more realistic estimates.

dimensionality. For scalability reasons, we linearly down-project the $cpc_{tf.idf}$ embedding from 140k to 768 dimensions when integrating it into the TMM model. We hypothesize that this dimensionality reduction is responsible for the performance drop. We conclude that the combination of all patent text field embeddings and cpc_{graph} is most effective for the target classification tasks across datasets.

5.4 Comparison with the Baselines

Table 5 shows that our best-performing approach (TMM + $e(t+a) \oplus e(cl) \oplus e(desc); cpc_{graph}$) outperforms the baselines in terms of macro- and micro-F1 across the three datasets. The SVM model by Benites et al. (2018) excels in terms of recall, but our method achieves a much higher precision and hence higher macro- and micro-F1 scores, especially for the gold-standard datasets InjVal and Rito. Note that the prediction threshold of the SVM model is optimized to maximize the macro-F1 on the dev split.

Comparing to the neural baseline, TMM with $e(t + a)$, which has recently reported state-of-the-art results on CPC classification (Pujari et al., 2021), we find that adding information from additional text fields and the CPC embeddings consistently improves performance.

Our analysis has shown that abstracts are often duplicated across patents (see Section 3.4). The problem aggravates when performing a PLS within a narrow field, e.g., around an invention. Therefore, using additional textual content fields is paramount.

In summary, our proposed approach consistently outperforms the baselines across three datasets both in terms of micro- and macro-F1 due to balanced precision and recall scores. We hence suggest that it provides a robust method that can be used as the basis for future work and for target classification tasks in real-world PLSs.

5.5 Minimum Training Instances

Motivated by the high cost of manual labeling by domain experts, we perform a study to determine the minimum number of training instances required for training a classification model that has an acceptable performance over unseen data.

Figure 4 shows macro-F1 and micro-F1 scores over different training sizes where the training instances were randomly sampled. Across datasets, we observe an acceptable micro-F1 performance with a training set size between 200 to 300 instances. On Rito and Atz near-optimal micro-F1 is achieved with a training set size of 200 instances. On the InjVal dataset, for a training set size of 300, a micro-F1 of around 70 is achieved compared to the maximum micro-F1 score of 84.3 with the complete dataset with 4.8k instances. These results illustrate that with as few as 200 instances, systems can be developed that already have significant value to patent professionals. However, the lower macro-F1 indicates insufficient performance for infrequent target labels. If these categories are of interest, further research is needed on how to ensure good performance with little training data and on integrating user feedback, e.g., via active learning.

6 Conclusions and Future Work

To foster research in the field of automating PLSs, we have introduced the new task of target label classification and released three real-world datasets. We have compared various neural and non-neural methods with different input representations covering the patents' texts and CPC information. As a result, we propose a competitive neural patent classification model, which leverages both patent-text and the CPC label information, and which shows robust performance across all three datasets. We found that an acceptable performance in terms of micro-F1 can be reached with only 200 to 300 training instances, demonstrating the practical applicability of the approach.

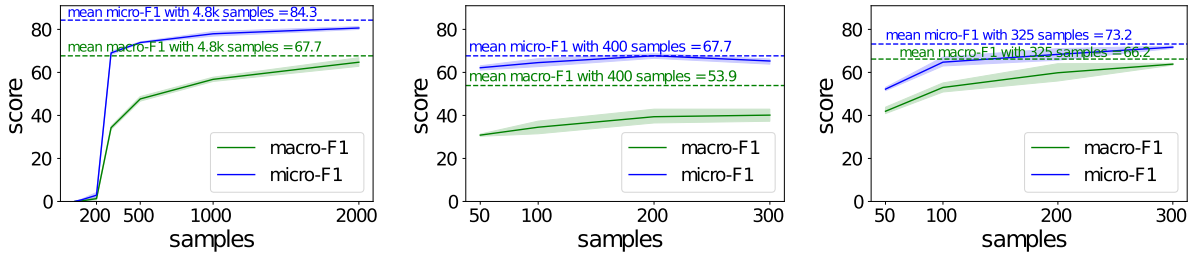
In order to improve performance for infrequent classes, integrating our methods with active learning of few-shot techniques are potential future directions. Our datasets also provide a valuable testbed for future work on neural representations for long and structured text documents.

Acknowledgements

We thank PatBase, MineSoft and RWS for giving us permission to release the patents along with their metadata and automatic translations. We also thank Ulrich Klingner for creating the annotations on the InjVal part of the dataset and for his valuable explanations. We also thank Irene Kitsara and Patrick Fievet from the World Intellectual Property Organization (WIPO) for an insightful discussion and information on the WIPO-related patent landscape studies. We also thank Tim Tarsi for fruitful discussions.

Dataset	Model	macro-avg. P	macro-avg. R	macro-avg. F1	micro-avg. P	micro-avg. R	micro-avg. F1
InjVal	Benites et al. (2018): SVM	64.0±3.8	**69.7**±7.1	61.4±2.1	61.7±7.0	**93.8**±2.3	74.1±4.3
InjVal	Pujari et al. (2021): TMM + $e(t+a)$	68.8±3.5	65.2±1.9	65.2±2.1	78.8±1.3	79.6±1.5	79.2±1.3
InjVal	TMM + $e(t+a) \oplus e(cl) \oplus e(desc)$; cpc_{graph}	**74.3**±6.7	66.4±1.6	**67.7**±2.5	**84.2**±0.8	84.4±0.6	**84.3**±0.5
Rito	Benites et al. (2018): SVM	46.6±13.0	**69.9**±4.6	51.1±6.8	43.1±3.3	**90.1**±2.3	58.2±2.5
Rito	Pujari et al. (2021): TMM + $e(t+a)$	58.5±5.0	42.2±4.6	44.3±4.0	67.8±2.9	64.3±3.9	66.0±3.0
Rito	TMM + $e(t+a) \oplus e(cl) \oplus e(desc)$; cpc_{graph}	**64.4**±6.9	49.5±6.5	**53.9**±6.6	**70.7**±2.6	65.1±4.9	**67.7**±3.2
Atz	Benites et al. (2018): SVM	66.7±7.8	**70.0**±5.9	65.4±1.8	65.2±6.4	**81.0**±4.4	71.9±2.5
Atz	Pujari et al. (2021): TMM + $e(t+a)$	68.6±1.4	59.7±4.0	62.1±2.2	73.0±2.6	68.8±4.4	70.6±1.1
Atz	TMM + $e(t+a) \oplus e(cl) \oplus e(desc)$; cpc_{graph}	**72.2±3.9**	63.3±1.5	**66.2**±1.8	**75.6**±4.1	70.9±1.9	**73.2**±2.1

Table 5: Comparison of our best-performing approach to the non-neural and neural baselines.



(a) InjVal (b) Ritonavir (c) Atazanavir

Figure 4: **Learning Curves**: performance for varying numbers of training instances.

Ethical Considerations

The datasets that we release with this paper are based on publicly available information. We ensured that they can be released under CC-BY 4.0 by (i) obtaining the explicit consent from the domain expert who has labeled the InjVal dataset, (ii) obtaining the explicit consent of WIPO to build on their data (which is already available under CC-BY 4.0), (ii) obtaining the permission from PatBase, RWS, and MineSoft to publish the metadata and translations obtained for the set of patents using their tools.

Limitations

Our dataset provides a benchmark for the representation and classification of long text documents. Our experiments show that relatively simple TF-IDF-based models perform competitively, but our study leaves computing results for long-range models such as LongFormer or BigBird to future work.

The dataset exists of English patents or patents translated into English; in future iterations, it may be highly interesting to construct multilingual patent landscaping dataset. Our dataset covers three patent landscape studies from two diverse domains. In the future, it would be desirable to add even more domains. The dataset provides an ideal testbed for methods addressing class imbalance and long-tailed settings. In its current form, the paper does not yet test such methods on the dataset.

References

Aaron Abood and Dave Feltenberger. 2018. Automated Patent Landscaping. *Artificial Intelligence Law*, 26(2):103–125.

Sophia Althammer, Mark Buckley, Sebastian Hofstätter, and Allan Hanbury. 2021. Linguistically Informed Masking for Representation Learning in the Patent Domain. In *Proceedings of the 2nd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech'21) co-located with the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'21)*, Online.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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'19)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Fernando Benites, Shervin Malmasi, and Marcos Zampieri. 2018. Classifying Patent Applications with Ensemble Methods. In *Proceedings of the 16th Annual Workshop of The Australasian Language*

Technology Association (ALTA'18), Dunedin, New Zealand.

Seokkyu Choi, Hyeonju Lee, Eunjeong Park, and Sungchul Choi. 2022. Deep Learning for Patent Landscaping Using Transformer and Graph Embedding. *Technological Forecasting and Social Change*, 175:121413.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning*, 20(3):273–297.

CSIRO's Data61. 2018. StellarGraph Machine Learning Library. https://github.com/stellargraph/stellargraph.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'19)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. 2003. Automated Categorization in the International Patent Classification. *SIGIR Forum*, 37(1):10–25.

Lintao Fang, Le Zhang, Han Wu, Tong Xu, Ding Zhou, and Enhong Chen. 2021. Patent2Vec: Multi-view Representation Learning on Patent-graphs for Patent Classification. *World Wide Web*, 24(5):1791–1812.

Alexander V Giczy, Nicholas A Pairolero, and Andrew A Toole. 2022. Identifying Artificial Intelligence (AI) Invention: A Novel AI Patent Dataset. *The Journal of Technology Transfer*, 47(2):476–505.

Mattyws F. Grawe, Claudia A. Martins, and Andreia G. Bonfante. 2017. Automated Patent Classification Using Word Embedding. In *Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA'17)*, pages 408–411, Cancun, Mexico. IEEE.

Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, page 855–864, New York, NY, USA. Association for Computing Machinery.

Jacques Guyot, Karim Benzineb, and Gilles Falquet. 2010. myClass: A Mature Tool for Patent Classification. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF'10)*, Padua, Italy. CEUR-WS.org.

Jieh-Sheng Lee and Jieh Hsiang. 2020. PatentBERT: Patent Classification with Fine-tuning a Pre-trained BERT Model. *World Patent Information*, 61(101965).

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.

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. *CoRR*, abs/1907.11692.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting Post Severity in Mental Health Forums. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 133–137. The Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Workshop Track Proceedings of the 1st International Conference on Learning Representations (ICLR'13)*, Scottsdale, Arizona, USA.

Muyao Niu and Jie Cai. 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)*, pages 3438–3443, Hong Kong, China.

Hyunji Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient Classification of Long Documents Using Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 702–709, Dublin, Ireland. Association for Computational Linguistics.

Subhash Chandra Pujari, Annemarie Friedrich, and Jannik Strötgen. 2021. A Multi-task Approach to Neural Multi-label Hierarchical Patent Classification Using Transformers. In *Proceedings of the 43rd European Conference on Information Retrieval (ECIR'21)*, volume 12656 of *Lecture Notes in Computer Science*, pages 513–528. Springer.

Subhash Chandra Pujari, Fryderyk Mantiuk, Mark Giereth, Jannik Strötgen, and Annemarie Friedrich. 2022. Evaluating Neural Multi-Field Document Representations for Patent Classification. In *Proceedings of the 12th International Workshop on Bibliometric-enhanced Information Retrieval (BIR'22) co-located with 44th European Conference on Information Retrieval (ECIR'22)*, volume 3230 of *CEUR Workshop Proceedings*, pages 13–27, Stavanger, Norway (hybrid). CEUR-WS.org.

Georg Richter and Andrew MacFarlane. 2005. The Impact of Metadata on the Accuracy of Automated Patent Classification. *World Patent Information*, 27(1):13–26.

Julian Risch, Nicolas Alder, Christoph Hewel, and Ralf Krestel. 2020. PatentMatch: A Dataset for Matching Patent Claims & Prior Art. *CoRR*, abs/2012.13919.

Julian Risch and Ralf Krestel. 2019. Domain-specific Word Embeddings for Patent Classification. *Data Technologies and Applications*, 53(1):108–122.

Benedek Rozemberczki and Rik Sarkar. 2018. Fast Sequence-Based Embedding with Diffusion Graphs. In *Complex Networks IX*, pages 99–107, Cham. Springer International Publishing.

Rossi Setchi and Irena Spasic. 2020. AI-assisted Patent Prior Art Searching-feasibility Study.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Harold Smith. 2002. Automation of Patent Classification. *World Patent Information*, 24(4):269–271.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith. 2017. Exploring the Use of Text Classification in the Legal Domain. In *Proceedings of the 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL'17) co-located with the 16th International Conference on Artificial Intelligence and Law (ICAIL'17)*, volume 2143 of *CEUR Workshop Proceedings*, London, UK. CEUR-WS.org.

Suzan Verberne and Eva D'hondt. 2011. Patent Classification Experiments with the Linguistic Classification System LCS in CLEF-IP 2011. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF'11)*, Amsterdam, The Netherlands. CEUR-WS.org.

Mihai Vlase, Dan Munteanu, and Adrian Istrate. 2012. Improvement of K-means Clustering Using Patents Metadata. In *Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM'12)*, volume 7376 of *Lecture Notes in Computer Science*, pages 293–305, Berlin, Germany. Springer.

Chih-Hung Wu, Yun Ken, and Tao Huang. 2010. Patent Classification System Using a New Hybrid Genetic Algorithm Support Vector Machine. *Applied Soft Computing*, 10(4):1164–1177.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'20)*, online.

Supplementary Material

A Corpus Statistics

In the following sections, we provide some additional statistics to highlight the characteristics of and the differences between the three datasets.

A.1 Publication Year

Figure 5 shows the time range for the three datasets InjVal, Rito, and Atz using the publication date of the earliest family member. The patents within the InjVal dataset are spread across approximately 100 years (1920 - 2019) with a majority of them being from the last 50 years. In contrast, the WIPO datasets have a narrow timeline of roughly 20 years (16 and 22 years for Atz and Rito, respectively). Because of the large time horizon, we can assume that the InjVal dataset has a higher topic drift than the other two datasets and also a more diverse language.

A.2 Patent Offices

An organization may file an invention across different patent offices around the globe to safeguard its business interests. However, these multiple filings are associated with a common patent family identifier. In Figure 6, we show the distribution of publications across different jurisdictions. It is interesting to note that most of the publications for the InjVal dataset are filed in Japan (JP) and Germany (DE), two of the primary hubs for industrial innovation. WIPO datasets have a higher number of worldwide filings (WO), followed by the United States (US) as the second most popular choice for filing a patent. This difference in the jurisdiction indicates the documents' language, where most of the documents within the InjVal dataset are in non-English language compared to WIPO datasets. Thus, the English texts in our datasets are partially machine-translated texts.

A.3 Problem of Duplicate Abstracts

During our analysis, we detected that abstracts across different patents might be identical. In particular, our analysis reveals that an abstract is often duplicated across patents, particularly those belonging to the same assignee, i.e., the organization filing a patent. As shown in Figure 7, the InjVal dataset contains seven abstracts that occur in at least two documents, whereas, in the case of Rito and Atz, the number of such abstracts is 20 and 25, respectively. For example, US7124963[10], US7137577B2[11], and US7198207B2[12] have identical abstracts, even though they belong to different patent families.

B Analyzing the Correspondence between CPC/IPC Labels and PLS-Oriented Target Labels

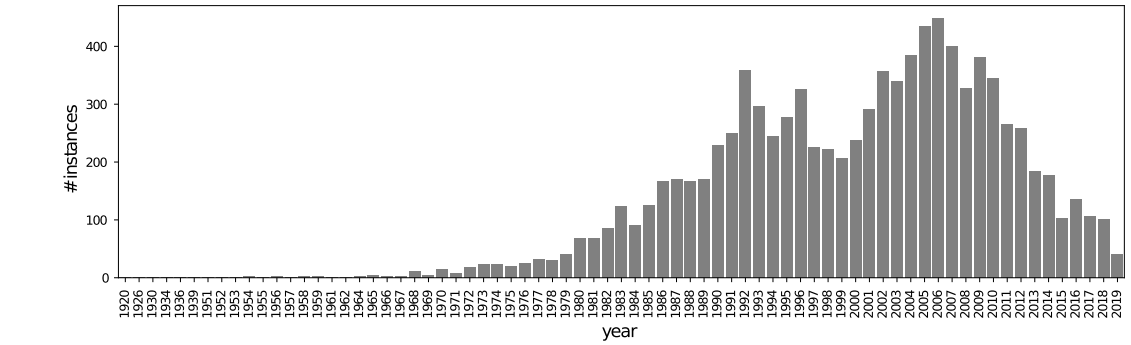
The value of CPC/IPC labels for the target classification depends on the correspondence between CPC/IPC labels and target labels. The hypothesis is that the higher the correspondence, the better the performance of a target classification method which exploits CPC/IPC information will be. We thus analyze the correspondence between CPC/IPC labels and target labels. We use Pointwise Mutual Information (PMI) (Church and Hanks, 1990) to measure correspondence. We calculate PMI between each CPC/IPC / target label pair and analyze it to determine the CPC/IPC and target label correspondence for each of the three datasets.

As a first analysis, we plot the PMI values of the top-50 CPC/IPC labels corresponding to the target labels. The underlying assumption is that if a CPC/IPC label is essential for a target label, it is also essential for the dataset. In Figure 8, we plot the PMI values for target labels and top-50 CPC/IPC labels for all three datasets. In InjVal (Figure 8a), a large number of CPC/IPC-target pairs have a very high PMI value, considerably higher than in the other two datasets. For Rito (Figure 8b), we identify fewer CPC/IPC-target pairs with high PMI values compared to Atz (Figure 8c).

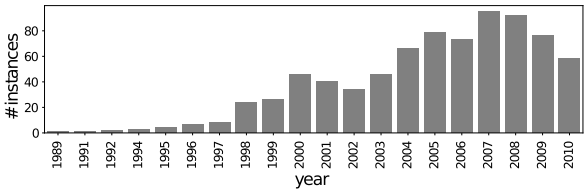
Further, to grasp the variation in PMI scores for top-k CPC/IPC labels, we plot the mean PMI value in Figure 9. The InjVal dataset shows a much higher mean PMI score across top-k CPC/IPC label counts compared to the WIPO datasets. Among the WIPO datasets, the mean PMI score is higher for Rito than Atz.

Summary. The main conclusions of our analysis can be summarized as follows: With our analysis, we find that the InjVal dataset shows a higher correlation between CPC/IPC and target labels compared to Rito and Atz. Among the WIPO datasets, Rito shows a higher correspondence than Atz.

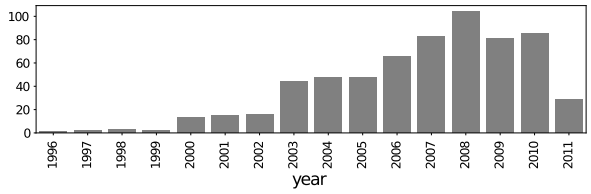
[10] https://patents.google.com/patent/US7124963B2
[11] https://patents.google.com/patent/US7137577B2
[12] https://patents.google.com/patent/US7198207B2



(a) Injection-Valve

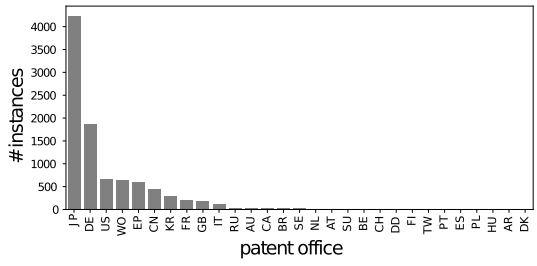


(b) Ritonavir

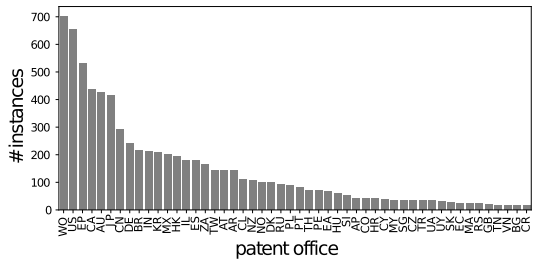


(c) Atazanavir

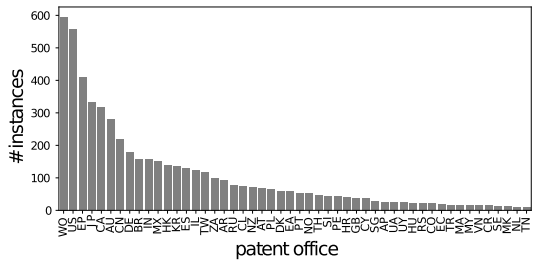
Figure 5: Instances per year. The InjVal dataset is from a much longer time horizon (around 100 years) when compared to Rito and Atz datasets.



(a) Injection-Valve



(b) Ritonavir



(c) Atazanavir

Figure 6: Patent office for different datasets.

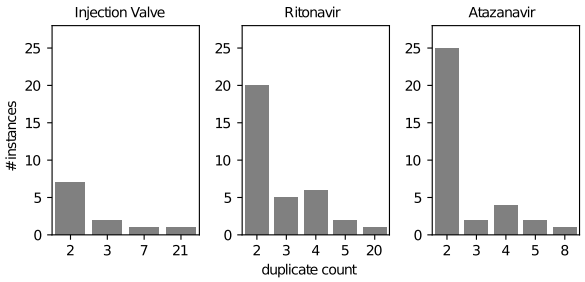
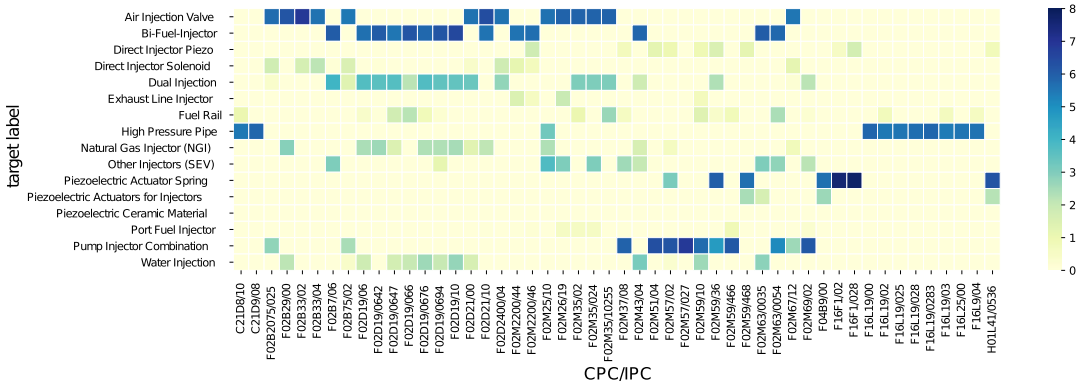


Figure 7: Duplicate abstract count.

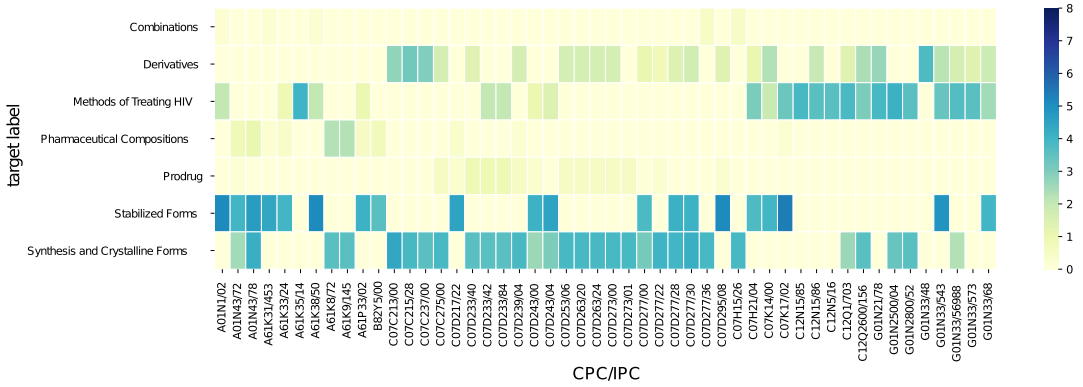
C Hyperparameters and Implementation

Below, we define the parameters for node2vec and the classification model. We use grid-search to identify the optimal hyperparamters for a classification model. The neural model experiments were performed on the Nvidia Tesla V100 GPU with 40GB VRAM.

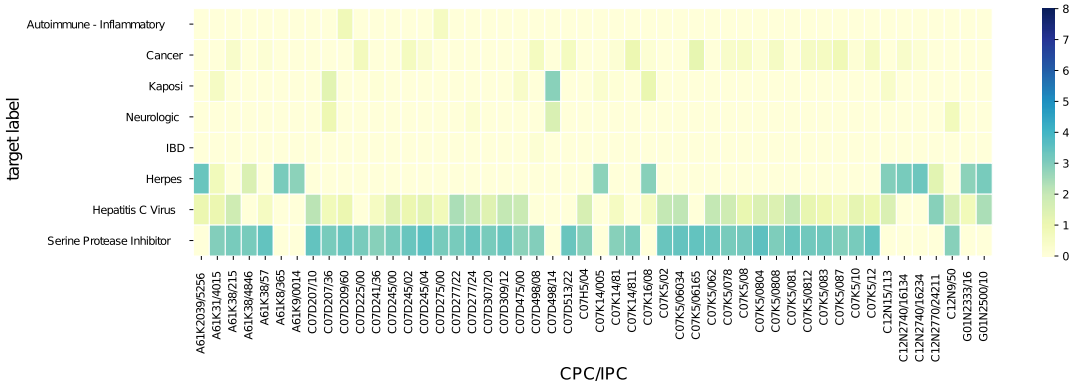
Node2Vec. Due to the large size of a global CPC co-occurrence graph (240k nodes and 40M edges), we consider a subgraph comprised of CPC labels from the three datasets with 9.5k nodes and 500k edges. This subgraph is provided as input to the node2vec algorithm generating the label-embeddings of dimension 128. The p and q parameters determine whether the next hop is selected from the neighbouring nodes or non-neighbouring nodes. Giving equal weightage to both these cases, we set the p and q values to 1. For computational



(a) InjVal



(b) Ritonavir



(c) Atazanavir

Figure 8: Plotting the correspondence between CPC/IPC and target labels.

efficiency, we perform 10 random walks with a maximum length of 50.

TMM. We use a hidden layer size of 50 for all dense layers in the classification heads, dropout set to 0.25 across layers, and a batch size of 4. We train all models for a maximum of 50 epochs with early stopping if the macro-F1 for the dev dataset does not increase for 7 epochs. We set a corpus-specific learning rate of 1e-05, 3e-05, and 5e-05 for InjVal, Rito, and Atz, respectively. The underlying SciBERT model is fine-tuned during training.

D Baseline with Target Label Names

Since we do not have exact details on the manual categorization process for the WIPO datasets, we experiment with a simple baseline searching for a label or associated keyphrases in the document text as documented in Table 7. Table 6 reports the results of using such a simple baseline for target classification. In general, we see that this simple baseline exploiting the label name and keyphrases results in a high recall on Atz and (though less high) Rito, precision is rather low, which shows the

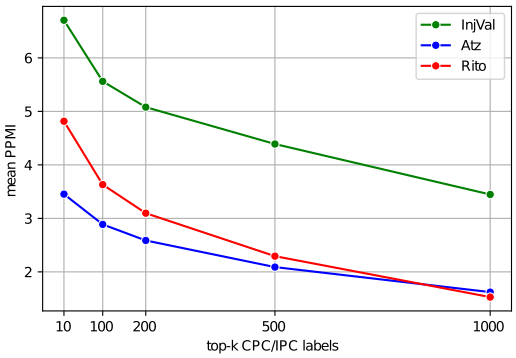


Figure 9: The plot shows the mean Pointwise Mutual Informatio (PMI) value of top-k labels. As we can see the InjVal dataset has much higher PMI values compared to the WIPO datasets. Among the WIPO datasets, Rito has higher PMI values than Atz.

need for a robust method to perform target label classification as suggested in Section 4. The simple keyword-based method does not work for the InjVal dataset due to extremely poor recall.

Based on our analysis, we conclude that for the Atz dataset, the domain expert extensively used the label name or associated keyphrases for determining document relevance. However, despite being supported by NLP technology as reported in the patent landscape report, we assume that the patent professional applied his or her domain expertise while labeling. For the InjVal dataset, we are aware that the domain expert primarily used the IPC codes and drawings within a patent document for judging the relevance.

Dataset	Model	macro-avg. P	macro-avg. R	macro-avg. F1	micro-avg. P	micro-avg. R	micro-avg. F1
InjVal	search with keyphrases on full-text	20.85	23.31	16.41	31.03	24.38	27.30
InjVal	search with label name on full-text	17.70	7.52	9.30	51.41	6.32	11.26
InjVal	search with label name on title + abstract	14.94	4.04	6.01	71.43	3.12	5.99
InjVal	our best (TMM + $e(t+a) \oplus e(cl) \oplus e(desc)$; cpc_{graph})	74.30	66.40	67.70	84.20	84.40	84.30
Rito	search with keyphrases on full-text	25.67	76.03	34.38	27.48	93.63	42.49
Rito	search with label name on full-text	23.39	42.24	28.24	36.52	68.15	47.56
Rito	search with label name on title + abstract	18.68	5.88	7.06	20.37	7.01	10.43
Rito	our best (TMM + $e(t+a) \oplus e(cl) \oplus e(desc)$; cpc_{graph})	64.40	49.50	53.90	70.70	65.10	67.70
Atz	search with keyphrases on full-text	49.11	86.51	56.71	42.53	83.58	56.38
Atz	search with label name on full-text	51.24	60.88	51.27	59.71	61.19	60.44
Atz	search with label name on title + abstract	55.21	11.05	17.11	89.29	12.44	21.83
Atz	our best (TMM + $e(t+a) \oplus e(cl) \oplus e(desc)$; cpc_{graph})	72.20	63.26	66.20	75.64	70.90	73.20

Table 6: Comparing the performance using simple baseline of searching label name or associated key phrases in the document text vs. our more sophisticated robust approach.

Dataset	Label	Keyphrases
InjVal	Exhaust Line Injector	exhaust line injector ; line injector
InjVal	Bi-Fuel-Injector	bi-fuel-injector ; bi-fuel injector
InjVal	Water Injection	water injection
InjVal	Piezoelectric Actuator Spring	piezoelectric actuator spring
InjVal	Fuel Rail	fuel rail
InjVal	Dual Injection	dual injection
InjVal	Direct Injector Piezo	direct injector piezo
InjVal	Port Fuel Injector	port fuel injector
InjVal	Direct Injector Solenoid	direct injector solenoid
InjVal	Air Injection Valve	air injection valve
InjVal	Piezoelectric Actuators for Injectors	piezoelectric actuators for Injectors ; piezoelectric actuator
InjVal	Pump Injector Combination	pump injector combination ; pump injector
InjVal	Other Injectors (SEV)	other injector
InjVal	Piezoelectric Ceramic Material	piezoelectric ceramic material ; piezoelectric ; ceramic
InjVal	Natural Gas Injector (NGI)	natural gas injector ; gas injector
InjVal	High Pressure Pipe	high pressure pipe ; high pressure
Rito	Pharmaceutical Compositions	pharmaceutical compositions ; composition ; pharmaceutical
Rito	Synthesis and Crystalline Forms	synthesis and crystalline forms ; crystalline form ; synthesis
Rito	Stabilized Forms	stabilized form
Rito	Methods of Treating HIV	methods of treating hiv ; hiv
Rito	Prodrug	prodrug
Rito	Derivatives	derivatives
Rito	Combinations	combination
Atz	Autoimmune - Inflammatory	autoimmune ; inflamatory ; autoimmune - inflammatory ; autoimmune-inflammatory ; autoimmune inflammatory
Atz	Cancer	cancer
Atz	Kaposi	kaposi
Atz	Neurologic	neurologic
Atz	IBD	ibd ; inflammatory ; bowel disease ; inflamatory bowel disease
Atz	Herpes	herpes
Atz	Hepatitis C Virus	hepatitis ; hepatitis c virus ; c virus
Atz	Serine Protease Inhibitor	serine protease inhibitor ; protease inhibitor ; inhibitor

Table 7: Keyphrases.