

# Twin (Green and Digital) Patents Identification: an Automated Patent Landscaping Method

Francesca Ghinami<sup>1,\*,†</sup>

<sup>1</sup>University of Cagliari, Department of Economics and Business, Via Sant'Ignazio 17, 09123, Italy

## Abstract

Identifying green, digital, and twin-transition patents is essential for tracking innovation and assessing policy impact, yet existing code-based and machine-learning approaches often yield non-overlapping results, undermining comparability and reproducibility. This study introduces a scalable framework that combines configurable keyword and technology rules for candidate identification, a rule-guided seed and antiseed definition, and bidirectional citation expansion. Patent texts are encoded with a domain-specific transformer, and final selection is achieved through topic-guided pruning based on a contrastive cosine rule applied to topic-level representations. Validation against proxy labels on a held-out split indicates high precision under a conservative threshold and balanced performance under a data-driven threshold. The workflow is automated, largely unsupervised, and tractable at the scale of millions of patent families, with results robust to sensible hyperparameter choices and threshold selection, thereby improving transparency and comparability for patent landscaping in the green and digital domains.

## Keywords

Patent Landscaping, Rule-based, Topic-guided pruning

## 1. Introduction

Identifying patents at the intersection of environmental and digital technologies (“twin patents”) is pivotal for tracking innovation dynamics and evaluating policies that foster sustainability and digitalization. Yet current approaches—ranging from examiner- or expert-selected technological codes and curated keywords, to citation-based heuristics and machine-learning pipelines—often select markedly different sets of documents [1], hampering comparability across studies and policy evaluations. Overlaps between sets built with different methods are very low—with observed Jaccard indices below random expectation, as reported in Table 3—underscoring the need for transparent, reproducible pipelines.

To address fragmentation while preserving transparency and scalability, this work introduces an automated workflow that minimizes manual intervention and combines rule-guided seed construction, two-level bidirectional citation expansion of the candidates, and topic-guided semantic pruning with patent-specific transformer embeddings.

First, candidate “twin” patents are identified by systematically crossing green selection strategies from the literature—namely the Cooperative Patent Classification (CPC) code *Y02* and targeted sustainability keywords—with digital strategies (CPC *Y04*, selected technological groups from the International Patent Classification (IPC), and digital/AI keywords). This *Candidates* set is then used to automatically derive three working sets: (i) the *seed* set, comprising high-precision exemplars of twin patents; a family enters the Seed if it is flagged by more than two independent modules (i.e., at least three; strict voting rule); (ii) the *expansion* set, obtained by expanding the Candidates via forward and backward family-level citations in two waves to collect plausibly related families; and (iii) the *antiseed* set, a size-matched control sampled outside both the Seed and the Expansion, designed to include mostly random non-twin patents and a 10% share of hard negatives. These hard-to-classify negatives are sampled from patents tagged as green or digital through *Y02* and *Y04* CPC codes, but not classified as jointly green-and-digital by

ITADATA2025: The 4<sup>th</sup> Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

\*Corresponding author.

✉ francesca.ghinami@unica.it (F. Ghinami)

🌐 <https://sites.google.com/view/francescaghinami/> (F. Ghinami)

🆔 0000-0003-1539-114X (F. Ghinami)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

any method (see Section 3.2). A patent-tailored encoder (PaECTER, [2]), built on Bidirectional Encoder Representations from Transformers (BERT), is used to obtain text embeddings, and a BERT-based topic model (BERTopic, [3]) projects documents into topic space. The expansion set is then pruned via a topic-guided criterion based on maximum cosine similarity to seed versus antiseed topics. Unsupervised diagnostics indicate clearer topic separation and a rightward shift in cosine-similarity densities toward seed topics after pruning. Finally, a pseudo-labeled evaluation set based on CPC Y02 $\cap$ Y04 codes—used as an expert-curated proxy for twin patents—is constructed for validation and for selecting a robust operating point by maximizing the Matthews correlation coefficient (MCC) on a stratified held-out split. On the same pseudo-labeled set, supervised-style metrics (precision, recall, F1, MCC) remain high under both a conservative threshold ( $\tau = 0$ ) and the MCC-maximizing threshold. A hyperparameter sensitivity study across the topic-discovery pipeline—including the text representation, low-dimensional projection, and density-based clustering stages—indicates that selection performance is insensitive to reasonable variation in these settings. At scale, the pipeline runs on PATSTAT Autumn 2024, leveraging structured metadata (technological codes, citations, abstracts and titles) from approximately 47M patent families with an English abstract. Simple patent families, grouping patent applications and publications sharing the same priority, are used as unit of analysis.

This article (i) proposes a reproducible, weakly-unsupervised seed definition procedure that integrates different sets of conditions derived from the literature, with a strict voting rule to balance breadth and precision; (ii) develops a citation-network aware expansion and matched random antiseed construction to enable unsupervised, contrastive pruning at scale; (iii) introduces a topic-guided semantic pruning strategy that couples domain-specific patent embeddings (PaECTER) with BERTopic topic distributions and a seed-vs-antiseed cosine decision rule, yielding cleaner, more coherent landscapes; and (iv) delivers a scalable, transparent pipeline with practical diagnostics and an accompanying Python implementation for researchers and policy analysts.

Section 2 reviews related methods and the overlap problem; Section 3 details seed rules, expansion, and topic-guided pruning; Section 4 describes data; Section 5 reports diagnostics; Section 5.5 presents the pseudo-label evaluation and robustness; Section 6 discusses limitations and future directions; Section 7 concludes.

## 2. Literature Review: Existing Methods, Potential and Limitations

The classification and identification of green and digital technologies have become central to understanding innovation dynamics in the context of the “twin transition”, which couples sustainability objectives with digital transformation [4]. Patents are widely used as proxies for innovation because they contain detailed technical descriptions and structured classification codes [5]. However, accurately identifying relevant patents—especially those that simultaneously address environmental and digital domains—remains methodologically challenging.

One group of approaches relies primarily on classification codes such as the International Patent Classification (IPC) or the Cooperative Patent Classification (CPC). These codes are assigned by examiners and enable systematic, replicable searches [6, 7]. While effective in principle, code-based methods often misalign with industrial categories and fragment technologies across classes, limiting their precision [5, 8].

Keyword-based searches offer a more flexible alternative, capable of capturing emerging or cross-cutting technologies [9, 1]. However, this flexibility introduces challenges, including linguistic variability, ambiguity, and potential biases in terminology that evolve over time or differ across jurisdictions [1]. Keyword methods also depend heavily on expert input to construct comprehensive queries, and can suffer from endogeneity when used in combination with machine learning [9].

Citation-based techniques leverage references among patents to identify related inventions and trace knowledge flows [10]. While citations can reveal important technological linkages, they are also shaped by examiner practices and strategic applicant behavior, leading to noise and incomplete coverage [11].

Recent work has increasingly combined these strategies. Integrated approaches, such as those

underpinning the IPC Green Inventory or ENV-TECH classification systems, attempt to blend the strengths of codes, keywords, and expert rules to improve recall and precision [12, 4]. However, evidence suggests that even these comprehensive frameworks often yield low overlap across methods, which limits their comparability and robustness [4].

Finally, advances in machine learning have introduced new possibilities for automating patent classification. Supervised and semi-supervised models trained on expert-labeled seed sets can extend coverage and reduce manual effort [10, 8]. Yet, the performance of these models depends critically on the quality and representativeness of the training data [13]. Further, computational resource requirements remain significant, particularly for models based on large transformer architectures such as BERT-Bidirectional Encoder Representations from Transformers [14] and its domain-adapted variants [2]; [15]; [16].

No single approach fully resolves coverage, accuracy, and scalability. This fragmentation, together with low overlap across methods, motivates integrated, open, and reproducible pipelines such as the one proposed here.

### 3. Methodology: towards an integrated approach

This study builds on prior semi-supervised patent landscaping methods [10, 8, 13, 2] and introduces several adaptations designed to improve replicability while minimizing human intervention. The proposed framework substitutes manual seed and antiseed selection with rule-based criteria, integrates bidirectional citation expansion, and applies transformer-based embeddings in combination with topic modeling and a pruning strategy based on the cosine similarity [17, 18] of topic-probability distributions. This design seeks to balance coverage, scalability, and semantic coherence in patent identification.

Earlier approaches typically relied on human-curated seeds and antiseeds [10], followed by expansions targeting overrepresented technological classes and citation networks. Subsequent classification models, trained on these curated examples, distinguished relevant patents based on semantic features. While effective, this process remained labor-intensive and sensitive to subjective decisions. To address these limitations, a generalizable rule-based approach for both seed and antiseed selection is adopted. A further improvement concerns the text classification model. Prior work often used static embeddings (e.g., Word2Vec), which struggled with ambiguity and polysemy. [8] tested alternative architectures—including MLP, CNN, and BERT—and identified BERT Transformers as the most consistent performers. Building on [2], the proposed method relies on PaECTER, a transformer optimized for patent texts and citations.

Finally, given the heterogeneity of twin patents combining digital and sustainability-related technologies, novel unsupervised methods are introduced to test the adherence of the expanded set to the seed. Specifically, BERTopic modeling [3] is applied to the seed, antiseed and expansion set texts (abstract and title), and expansion candidates are pruned based on cosine similarity to topic distribution vectors. The following sections describe each step in detail, while an overview of the method is shown in Figure 1. The replication code to run this pipeline is available at <https://github.com/GhinamiF/TwinPatentLandscape>, release v1.0.

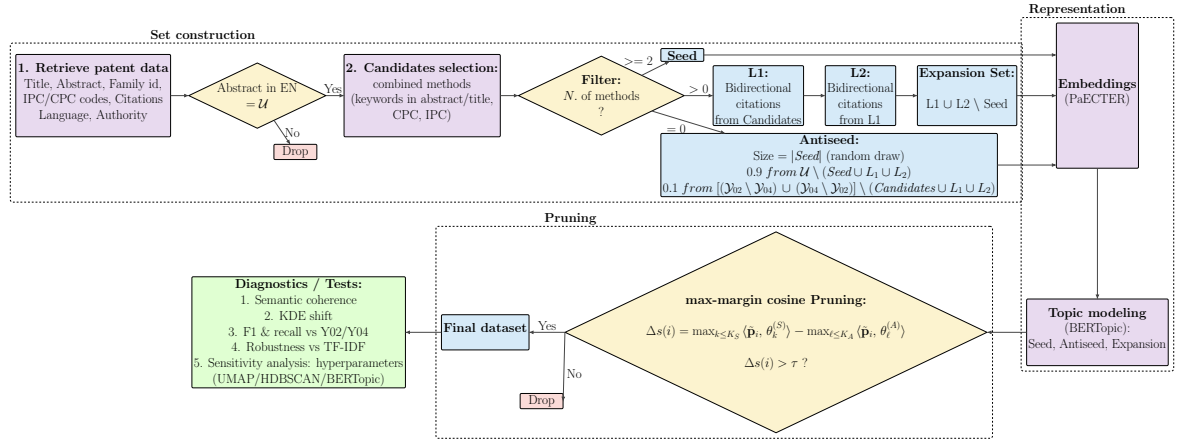
#### 3.1. Rule based seed selection

The seed set forms the foundation of the patent landscape, making its composition critical to ensure both accuracy and representativeness. To reduce reliance on manual curation and enhance replicability, a rule-based approach inspired by the strategy proposed by [4] is applied.

In their method, patents are identified as “twin” if they combine digital and green characteristics, based on CPC codes and keyword presence. Specifically, they apply six selection rules that combine Y02 (green technologies) and Y04 CPC codes (digital technologies), with keywords. Additionally, IPC codes are used to capture relevant groups and subclasses. Given the comprehensive nature of this framework<sup>1</sup>, their

---

<sup>1</sup>Details of the keyword and code lists are reported in Table 1.



**Figure 1:** Flow chart of the proposed method: rule-based seed, two-level bidirectional citation expansion, random and augmented antiseed, topic modeling, topic-guided pruning by cosine similarity, and testing.

method is adopted and extended to allow broader generalization to other technological domains. Unlike the original formulation, which applied a fixed set of combinations, here an adaptation is suggested to systematically combine any green identification strategy with any digital strategy in all possible pairings. This means that every sustainability-related rule (e.g., Y02 codes or green keywords) was crossed with every digital or AI-related rule (e.g., Y04 codes, digital IPC codes, or digital keywords), generating an expanded and more granular set of inclusion criteria. This ensures a balanced and comprehensive coverage of patents that may reflect diverse configurations of digital and green technologies. A patent was thus included in the initial candidate pool if it matched any of these combined rules. To improve precision and avoid reliance on any single identification method, a stringent inclusion criterion is applied: only patents identified as twin by more than two of the resulting combinations were retained in the seed set. This procedure yielded a final set of 9,847 unique patent families. These patents were considered sufficiently diverse and representative of the target technological intersection for subsequent expansion and pruning phases.

### 3.2. Expansion

The expansion methodology implemented in this study largely follows the approach outlined by Abood and Feltenberger (2018), which involves a two-tiered expansion process. However, the first level of expansion based on the most relevant CPC codes is here excluded, to mitigate the risk of over-relying on this information, as CPC codes are already incorporated as a rule in the definition of seed patents.

The first level of expansion (Level 1) involves identifying patents related to seed patents through backward and forward citations. This level includes all patents that cite the seed patents or are cited by them. Moreover, it is augmented by all the patents identified by any of the candidate selection method, but not included in the seed. The second level of expansion (Level 2) further extends this network by including patents that are related to the patents identified in Level 1 through their own backward and forward citations.

The antiseed set, serving as negative examples for later semantic comparison, was generated by randomly sampling an equal number of patent families not included in either the seed or expansion sets.

### 3.3. Transformer-based embeddings

In patent classification and pruning, models can be trained from scratch or adapted from domain-specific encoders. Training from scratch offers flexibility but typically requires substantial compute. A practical alternative is a pre-trained model tailored to patents. This study uses the PaECTER encoder [2], as

**Table 1**

Keywords, CPC, and IPCs codes used in the seed twin patents identification step [4].

Digital keywords/AI keywords	3D print, adaptive robotic, augmented reality, autonomous, big data, blockchain, business analytic, chip technology, cloud comput, cyber, data analytic, data transmission, data-based, digital, digitization, distributed comput, ebanking, e-commerce, ehealth, elearning, e-banking, e-commerce, e-health, e-learning, fog comput, industry 4.0, information system, intelligence, intelligent, internet, machine learning, mobile comput, natural language processing, quantum comput, smart, software, traffic optimi, virtual adaboost, artific intelligen, bayes network, bayesian belief networks, bayesian-network, bio-inspired approach, bio-inspired comput, biologically inspired comput, chatbot, classification tree, collaborative systems, computation intelligen, computer vision, connectionis, crowdsourcing and human computation, data mining, decision making, decision model, decision tree, Decision tree learn, deep learn, deep structured learn, description logistic, expert system, fuzzy logic, genetic algorithm, gradient tree boosting, graphical model, hidden markov model, hierarchical learn, humanoid robotics, image alignment, image grammars, image matching, inductive logic programm, instance-based learn, knowledge representation and reasoning, latent dirichlet allocation, latent represent, latent semantic analysis, layered control systems, learning algorithm, learning model, Logic Programming, logic theorist, logical learn, logistic regression, machine intelligen, machine learn, memory-based learn, multi-agent system, multilayer perceptron, multitask learn, natural language, neural network, neuromorphic computing, ontology engineer, optimal search, pattern analysis, pattern recognition, physical symbol system, probabilistic graphical model, probabilistic reason, probabil logic, random forest, rankboost, regression tree, reinforcement learn, relational learn, robot, robot systems, Rule induction, rule learn, semi-supervised learn, semi-supervised train, sensor data fusion, sensor network, stochastic gradient descent, structured probabilistic model, supervised learn, supervised train, support vector machin, support vector network, swarm intelligen, symbol processing, symbolic error analysis, symbolic reasoning, systems and control theory, task learn, transfer learning, xgboost
Green Keywords	air polluti, biodiversity, biofuel, carbon footprint, circular economy, clean energy, clean fuel, climate change, climate disaster, CO2 emission, CO2 level, eco-friend, electric vehicle, energy consumption reduc, environmental protect, environmental-friendly, environmentally-friendly, food waste, pollution control, pollution detect, GHG reduc, green energy, recycling, renewable energy, resource efficien, resource-efficien, green cit, smart farming, solid waste, sustainability, waste management, water efficiency, water leakage, water management, water scarcity, water treatment, greenhouse gas, reduction, carbon emission, pollution, resource consumption
Digital CPC codes	Y04
Green CPC codes	Y02
Digital IPC 5 digits codes (groups)	G02F 7, H03M 1, H03H 17, G06J 3, G06T 19, G06J 1, B33Y 50, G05B 15, G16H 80, G08C 17, G16H 40, H02J 13, H03L 9, G08B 19, H04H 60, G05B 17, G09C 5, F24F 130, G16H 20, G16H 15, G06F 9, G06F 8, G16H 50, G06T 17, H03H 21, H04L 12, G06F 21, F24F 120, G01R 13, G06Q 30, H04L 9, G06T 13, G06Q 20, G16H 30, H04L 1, G06Q 50, G09B 9, H04W 12, G06Q 10, G06T 15, G05D 1, G06F 15, G06E 1

BERT-based encoders have shown strong and consistent performance in patent analytics [8], and PaECTER, in particular, is reported as a top-performing patent-specific transformer [2]. As described in [2], PaECTER is trained on a patent-focused vocabulary and fine-tuned using a citation graph over a large corpus of English-language patent families (PATSTAT 2023 Spring).

Documents in the *Seed*, *Antiseed*, and *Expansion* sets—formed by concatenating title and abstract<sup>2</sup>—are

<sup>2</sup>All numerals are removed and text is lowercased.



encoded jointly with PaECTER to obtain a shared embedding space that enables comparable topic modeling and pruning. They are then organized with BERTopic [3], which combines UMAP [19] and HDBSCAN [20, 21] to produce a topic-probability vector  $p_i \in [0, 1]^K$  for each document. As a starting point, commonly used defaults are adopted—UMAP  $n_{\text{neighbors}}=15$ ,  $\text{min\_dist}=0.1$ ,  $n_{\text{components}}=5$ ; HDBSCAN  $\text{min\_cluster\_size}=10$ ,  $\text{min\_samples}=\text{None}$ ; BERTopic’s default vectorizer—and robustness is assessed in sensitivity checks.

The BERTopic mapping serves two roles: (i) it enables unsupervised diagnostics via topic-level metrics (Section 3.5); and (ii) it provides the representation on which the topic-guided pruning operates (Section 3.4 below).

### 3.4. Topic-guided pruning via a multi-prototype contrast

Prior work [8, 10] commonly prunes expansion sets with a global cosine-similarity rule (e.g., keep items sufficiently close to a seed centroid). For twin-transition patents—topically heterogeneous and lexically overlapping with near domains—such global rules can drop legitimate variants and retain marginal cases; empirically, seed and antiseed embedding distributions are only weakly separated.

To select relevant documents from a large, noisy expansion set in an unsupervised manner, we apply a *topic-guided pruning* strategy inspired by weak supervision and semi-supervised learning [22, 23, 24]. Two reference groups are considered: seeds  $\mathcal{S}$  (twin exemplars) and antiseeds  $\mathcal{A}$ . To capture sub-themes, multiple *prototypes* are learned by clustering seed and antiseed topic vectors into  $K_S$  and  $K_A$  groups (MiniBatch  $k$ -means [25, 26]), yielding L2-normalized centers  $\Theta_S = \{\theta_1^{(S)}, \dots, \theta_{K_S}^{(S)}\}$  and  $\Theta_A = \{\theta_1^{(A)}, \dots, \theta_{K_A}^{(A)}\}$ . Each document  $d_i$  is represented by its topic-probability vector  $\mathbf{p}_i$ ; let  $\tilde{\mathbf{p}}_i = \mathbf{p}_i / \|\mathbf{p}_i\|_2$  denote its L2-normalized version so that cosine similarity reduces to a dot product.

The pruning score is a *best-seed vs. best-antiseed* cosine margin:

$$\Delta s(i) = \underbrace{\max_{k \leq K_S} \langle \tilde{\mathbf{p}}_i, \theta_k^{(S)} \rangle}_{\text{closest seed prototype}} - \underbrace{\max_{\ell \leq K_A} \langle \tilde{\mathbf{p}}_i, \theta_\ell^{(A)} \rangle}_{\text{closest antiseed prototype}}.$$

A document is retained when  $\Delta s(i) \geq \tau$ . The primary operating point on the unlabeled corpus is the *zero-contrast* rule ( $\tau=0$ ), which keeps a document when it is at least as similar to seed prototypes as to antiseed prototypes. This choice is parameter-light, interpretable, and does not require labels.

### 3.5. Evaluation: semantic coherence, separation, and dispersion

We evaluate pruning within the BERTopic framework using established topic-model diagnostics. First, an intertopic distance map projects topic representations into two dimensions to visualize distinctiveness and dispersion, assessing whether pruning improves semantic coherence and separation [27, 3]. Second, to assess document-level alignment, we plot kernel density estimates of cosine similarity between expansion documents with respect to seed and antiseed prototypes in topic space, examining whether pruning increases alignment with seeds and reduces overlap with antiseeds [28, 29]. Together, these analyses provide global (topic structure) and local (document–prototype alignment) perspectives on pruning quality in high-dimensional, embedding-based topic models.

This approach combines the high recall of unsupervised expansion with the precision gains of topic-guided selection, and builds on guided topic discovery [23], weakly supervised labeling [24], and zero-/few-shot cross-domain classification [22].

### 3.6. Validation and robustness

In the absence of costly and hard-to-replicate human annotation, we validate our topic-guided pruning against a proxy gold standard derived from CPC Y02/Y04 codes. These codes are curated jointly by two patent authorities, the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO), and assigned by trained examiners, providing a practical proxy for human labels.

The evaluation pool comprises 29,032 patent families tagged with both Y02 and Y04 (treated as positives) and 9,847 antiseed families (treated as negatives, by construction a mix of random non-twins and hard negatives). We report standard retrieval metrics: precision (the fraction of selected patents that are true twins), recall (the fraction of true twins that are selected), and F1 (the harmonic mean of precision and recall, summarizing the precision–recall trade-off).

The evaluation pool also allows us to fine-tune and test different thresholds for pruning.

**Threshold selection on a pseudo-labeled evaluation set.** For quantitative assessment, a pseudo-labeled *evaluation* pool is built as  $Y02 \cap Y04$  CPC families (positives) together with antiseeds (negatives). A single 75/25 stratified train–test split is drawn from the pseudo-labeled evaluation pool. Thresholds are selected by maximizing MCC on the train split; all metrics are reported on the held-out test split.

On the training split, an MCC-optimal operating point is selected by

$$\tau_{\text{MCC}}^{\Delta s} \in \arg \max_{\tau} \text{MCC}(y, \mathbb{1}\{\Delta s \geq \tau\}),$$

and all supervised-style metrics (precision, recall, F1, MCC, accuracy, Jaccard and the confusion matrix) are reported for both  $\tau=0$  and  $\tau_{\text{MCC}}^{\Delta s}$ .

**Baseline for comparison (TF–IDF margin).** As a robustness check, a sparse lexical baseline is included that operates at the individual-document level in TF–IDF space [30]. Let  $\mathbf{x}_d$  be the L2-normalized TF–IDF vector of document  $d$  (built on  $\text{seeds} \cup \text{antiseeds}$ ). The TF–IDF margin scorer is

$$m(d) = \underbrace{\max_{s \in \mathcal{S}} \cos(\mathbf{x}_d, \mathbf{x}_s)}_{\text{closest seed doc}} - \underbrace{\max_{a \in \mathcal{A}} \cos(\mathbf{x}_d, \mathbf{x}_a)}_{\text{closest antiseed doc}}, \quad \text{retain if } m(d) \geq \tau.$$

On the same training split,  $\tau_{\text{MCC}}^{\text{TF-IDF}}$  is obtained by maximizing MCC, and test metrics are reported at  $\tau=0$  and  $\tau_{\text{MCC}}^{\text{TF-IDF}}$ . Because TF–IDF compares documents directly at the lexical level, it commonly achieves slightly higher supervised metrics on the evaluation split; however,  $\Delta s$  remains the preferred primary scorer for the full unlabeled corpus, as it enables unsupervised diagnostics, topic-level interpretability, and a principled  $\tau=0$  operating point independent of labels, with both  $\tau_{\text{MCC}}^{\Delta s}$  and the TF–IDF baseline serving as robustness checks.

## 4. Data

The empirical analysis is conducted using the Patstat Autumn 2024 database [31], a comprehensive dataset of global patent records maintained by the European Patent Office (EPO). This database includes detailed bibliographic information, classification codes, citation linkages, and legal status for millions of patent families worldwide. The database comprises 85,195,446 patent applications grouped into 66,798,016 DOCDB simple families, 47,068,344 of which include an English abstract. We restrict our analysis to these families to enable text-based semantic analysis. In Patstat 2024 [32], every patent application is assigned to a simple family, also known as the DOCDB family, which links applications that share exactly the same priority [32]. This differs from the extended family (INPADOC family), which links applications sharing a priority either directly or indirectly through a third application. The choice of using the simple family as unit of analysis offers several advantages. It avoids the issue of over-representation of seed inventions that are published under multiple IDs in different jurisdictions, ensuring a more accurate representation of the related patents in the expansion set. Additionally, it allows for the identification of all citation linkages, ensuring that the citation network is fully captured. According to the EPO’s Data Catalog [32], simple family citations encompass both citations to patent publications and applications.

An example of the data collected is shown in Table 2.

**Table 2**

Example of collected Patstat data, for a patent drawn from the seed set.

Field	Value
<b>Title</b> (title_text)	CHARGING STATION MONITORING SYSTEM
<b>Abstract</b> (abstract_text)	A charging station monitoring system comprising: a sensing device, a digital camera and a communication device which are arranged at a charging apparatus of a charging station, the sensing device and the digital camera each having a sensing range covering a parking lot associated with the charging apparatus and an area around the parking lot; and a controller configured to determine an occupation state of the parking lot and/or detect and record an action of a third party or foreign object based on sensed information from the sensing device and the digital camera.
<b>Application id</b> (appln_id)	521014361
<b>Publication id</b> (pat_publn_id)	530024524
<b>Simple family id</b> (docdb_family_id)	68461712
<b>Filing date</b> (earliest_filing_date)	2018-11-20
<b>Citations</b> (cited_docdb_family_id)	43031012, 46705595, 51896513, 53544075, 54767462, 55201075, 55396976, 56309691, 57758888, 57906422, 59333789, 61477574, 62481294, 63917493
<b>CPC codes</b> (cpc_class_symbol)	B60L 53/30, B60L 53/60, B60L 53/68, G06V 10/95, G06V 20/52, H01M 10/44, H02J 7/0013, H02J 7/007192, H04N 7/18, Y02E 60/10, Y02T 10/70, Y02T 10/7072, Y02T 90/12, Y02T 90/16, Y02T 90/167, Y04S 30/12
<b>IPC codes</b> (ipc_class_symbol)	B60L 53/30, B60L 53/60, B60L 53/68, G06K 9/00, G06T 1/00

## 5. Results

### 5.1. Candidates selection methods

As an initial analysis, the identification strategies proposed by [4] are replicated. Using their modules, it is possible to classify a total of 238,717 patent families as ‘twin’ by at least one module. However, as it can be shown, there is minimal to no overlap among the patent sets identified by their different methods. This highlights a challenge regarding the representativeness of the resulting seed set. To address this, the strategies detailed in Section 3 are proposed, which combine the same set of information, reported in Table 1, using alternative configurations. As reported in Table 3 This procedure yields 223,575 unique patent families classified as twin by at least one method. The low overlap across methods remains, as shown by the average Jaccard metric computed for all methods (0.064) which is lower than the random assignment one (obtained through same set sizes and 3,000 repetitions).

### 5.2. Seed, anti-seed and expansion set analysis

Applying the most stringent criterion —requiring identification by more than two methods to be included in the seed— results in a final set of 9,847 unique seed patent families, which can be considered as the most representative and thus suitable for a robust seed definition. By expanding the candidates set twice through bidirectional citations, an expansion set comprising 1,918,509 unique patent families is derived.

### 5.3. Topic modeling results

After preprocessing (lowercasing and removing numerals), the seed, antiseed, and expansion documents are encoded and modeled with BERTopic. The model yields 86 topics for the overall set (Figure 2). The intertopic distance visualization (Figure 2, panel (a)) provides an overview of the topics generated by the BERTopic model for the overall set of seed, expansion and antiseed documents. Each point on the visualization represents a topic, with its position determined by the model’s learned embeddings. Topics that are close to one another on the map suggest similar themes, while those that are farther apart indicate distinct conceptual areas. This visualization revealed a well-structured topic space, with several clusters of topics grouped around core themes relevant to the twin patent corpus.

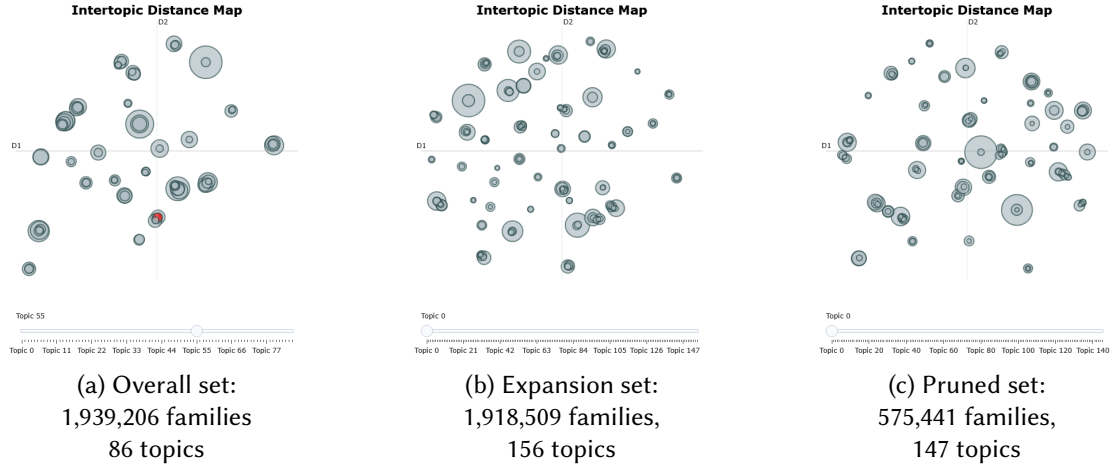
Overall, the BERTopic model’s initial output provides a coherent representation of the twin patent data’s thematic landscape. The visualizations demonstrates a structured topic space with clear thematic areas and revealed areas for further refinement in the subsequent pruning phase, where the expansion set would be filtered based on cosine similarity to seed topics.



**Table 3**

Twin patents identification methods revised.

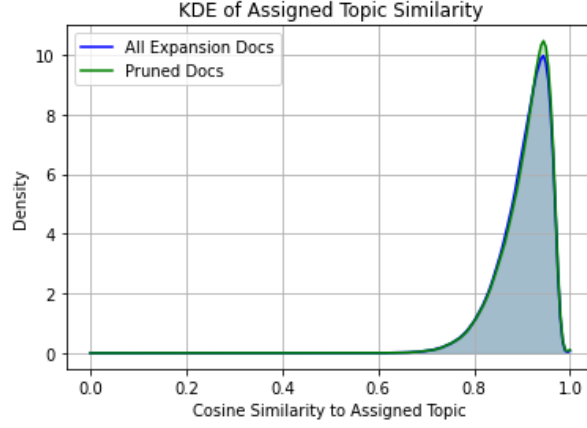
M	Description of identification	digital/AI	green	N. patent families	$J^{M1}$
1	Patents tagged with both Y04 and Y02	CPC(Y04)	CPC(Y02)	38,519	-
2	Patents with at least one digital+AI and one sustainability keyword in title or abstract	Keywords	Keywords	37,764	0.012
3	Patents with at least one digital/AI keyword in title or abstract and classified under Y02 code	Keywords	CPC(Y02)	112,435	0.078
4	Patents tagged with CPC(Y04) and at least one sustainability keyword in title or abstract	CPC(Y04)	Keywords	4,203	0.099
5	Patents classified in at least one digital IPC group (5 digits) and under Y02 code	IPC group	CPC(Y02)	87,311	0.249
6	Patents classified in at least one considered digital IPC group (5 digits) and containing one sustainability-related keyword	IPC group	Keywords	15,234	0.024
#	Total (unique, in english) Jaccard: $\bar{J}_{obs}=0.064$ ; $\bar{J}_{rand}=0.083$			223,575	

**Figure 2:** Intertopic distance graphs

## 5.4. Pruning results

By applying the pruning strategy described in Section 3.3 to the expansion set, 575,441 patent families are retained.

To assess the impact of the pruning strategy in an unsupervised setting, *id est* in absence of labeling that hinders the use of recall and precision analysis, a number of tests can be performed. First, the topic analysis on the whole and pruned expansion set can be repeated, in order to compare the Intertopic Distance Map [27] before and after pruning. Pre-pruning, topics appeared more crowded and overlapped substantially, suggesting semantic redundancy and poor topic separation. After pruning, the 2D projection revealed clearer topic dispersion, with reduced overlap and tighter clustering. This suggests increased semantic distinctiveness and topical coherence, both indicators of a better-defined topic space [3]. Highlighting this visually helps validate the pruning process not just by number reduction but by structural improvement in the latent topic space. After pruning, topics appear tighter and more coherent and, while more numerous than in the overall set (Figure 2), they are fewer than in the expansion set (147 vs 156), indicating that pruning removes broader, disjointed topics while preserving fine-grained themes.



**Figure 3:** Cosine Similarity to Assigned Seed Topics, for documents in the original and pruned expansion sets

**Table 4**

Test performance at  $\tau=0$  and MCC-optimal thresholds learned on the training split, separately for the multi-prototype margin  $\Delta s$  and the TF-IDF margin.

Scorer	Threshold	P	R	F1	Acc.	Jac.
$\Delta s$ (multi-proto)	$\tau=0$	0.973	0.888	0.928	0.900	0.866
$\Delta s$ (multi-proto)	$\tau_{MCC}^{\Delta s}$	0.932	0.981	0.956	0.934	0.915
TF-IDF	$\tau=0$	1.000	0.878	0.935	0.911	0.878
TF-IDF	$\tau_{MCC}^{TF-IDF}$	0.958	1.000	0.978	0.968	0.958

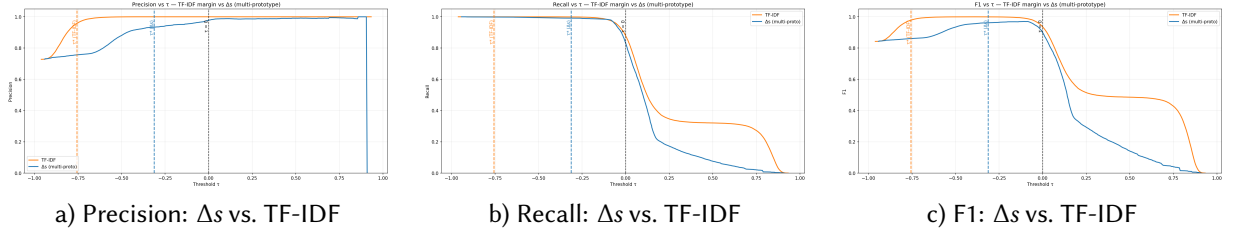
Finally, comparing the panels (b) and (c) in Figure 2, a greater average distance is found in the expansion set compared to the pruned one, and that broader and disjointed topics were effectively dropped.

Subsequently, to evaluate the semantic alignment of the expansion set with seed and antiseed topics, the Kernel Density Estimation (KDE) plots of the maximum cosine similarity between each expansion document and (i) seed topics and (ii) antiseed topics are produced. These plots provide a smooth estimation of similarity density across the corpus [28]. As shown in Figure 3, after pruning, the KDE curve shifted toward higher similarity with seed topics and away from antiseeds, indicating that retained documents are more aligned with the intended thematic focus and less with undesired content. This is consistent with effective filtering in vector space, aligning with known techniques in bias detection and semantic drift analysis [29].

These diagnostics confirm that pruning was both selective and semantically discriminative, reducing noise and enhancing alignment with the seed topics.

## 5.5. Validation against CPC pseudo-labels

In the first two rows of Table 4 we report validation of the baseline method—the multi-prototype topic contrast  $\Delta s$ —at the conservative threshold  $\tau=0$  and at a data-driven threshold  $\tau_{MCC}^{\Delta s}$  selected by maximizing the Matthews correlation coefficient on a separate development split of the labeled pool. At the conservative operating point ( $\tau=0$ ),  $\Delta s$  attains Precision=0.973, F1=0.928, and Jaccard=0.866. Using the data-driven threshold,  $\Delta s$  improves to F1=0.956 with higher recall (Recall=0.981) and slightly lower precision (Precision=0.932). This operating point typically increases recall (and overall F1) while allowing a controlled rise in false positives. Although the tuned threshold is favored in terms of precision, accuracy, and overlap, the conservative rule performs well and remains attractive when labels are unavailable. Practitioners can choose the operating point that matches their objective: the label-free  $\tau=0$  rule when high precision and simplicity are paramount, or the MCC-selected threshold when broader coverage (higher recall) is preferred, and labels are available.



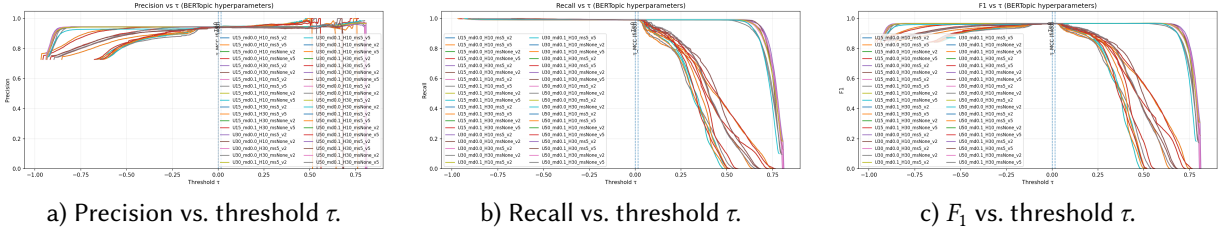
**Figure 4:** Precision, recall and F1 as functions of the pruning threshold  $\tau$  for two scorers: a topic-space multi-prototype contrast  $\Delta s$  (blue) a TF-IDF max-margin (orange). Vertical dashed lines mark the conservative rule  $\tau$  (black) and the data-driven thresholds ( $\tau_{MCC}^{\Delta s}$  and  $\tau_{MCC}^{TF-IDF}$ ), selected separately for each scorer (colored).

The same table also reports test performance for a TF-IDF max-margin variant evaluated at  $\tau=0$  and at its own TF-IDF-specific MCC-selected threshold.  $\Delta s$  operates in topic space to provide interpretable, prototype-aligned pruning, whereas TF-IDF offers a document-space robustness baseline. At  $\tau=0$ ,  $\Delta s$  achieves high precision with slightly higher recall than TF-IDF ( $F1 = 0.928$  vs.  $0.935$ ). With a data-driven threshold,  $\Delta s$  improves to  $F1 = 0.956$  with a strong recall gain, while TF-IDF reaches a higher balanced score overall ( $F1 = 0.978$ ).

Despite TF-IDF’s gain on this proxy-labeled test, we retain  $\Delta s$  as the primary pruning rule because it (i) operates in the same topic space that underpins our diagnostics, yielding interpretable prototype-level decisions; (ii) reduces lexical leakage from seed phrasing and is less sensitive to vectorizer settings and vocabulary drift; and (iii) provides a transparent, label-free  $\tau=0$  policy that performs well on the unlabeled corpus. We therefore report both scorers—using  $\Delta s$  for the main selection and TF-IDF as a robustness check—and include threshold-sensitivity curves in Figure 4 to make the trade-offs explicit. Both scorers display the expected trade-off: as  $\tau$  increases, precision rises while recall falls, and F1 peaks on a broad plateau. The TF-IDF margin reaches the highest peak F1 and attains near-perfect precision at more conservative  $\tau$ , dominating the upper-right region of the curves. The multi-prototype  $\Delta s$  increases precision more gradually and preserves relatively higher recall near  $\tau=0$ , providing a smooth, interpretable operating range aligned with the topic-space diagnostics. In practice, TF-IDF at its data-driven  $\tau_{MCC}^{TF-IDF}$  is preferable for balanced performance when labels (or proxy labels) are available, whereas  $\Delta s$  near  $\tau=0$  is a well-behaved default for label-free, topic-aligned pruning. Both methods are stable over a wide plateau of  $\tau$  values, so small threshold shifts do not materially affect results.

## 5.6. Sensitivity analysis across BERTopic hyperparameters .

To assess how the pruning threshold  $\tau$  in the topic-space contrast score  $\Delta s = \cos(\mathbf{p}, \mathbf{\hat{s}}) - \cos(\mathbf{p}, \mathbf{\hat{a}})$  affects retrieval quality—and how this behavior varies with modeling choices—the topic model is refit over a grid spanning the three stages of the pipeline: (i) *Text vectorization* (bag-of-words with  $n$ -grams), controlling vocabulary granularity via  $\text{min\_df} \in \{2, 5\}$ ; (ii) *Low-dimensional projection* (UMAP), controlling local neighborhood size and layout via  $n_{\text{neighbors}} \in \{15, 30, 50\}$  and  $\text{min\_dist} \in \{0.0, 0.1\}$ ; (iii) *Density-based clustering* (HDBSCAN), controlling cluster granularity and treatment of noise via  $\text{min\_cluster\_size} \in \{10, 30\}$  and  $\text{min\_samples} \in \{\text{None}, 5\}$ . For each configuration, topic probabilities are obtained, the contrast score  $\Delta s$  is recomputed, and  $\tau$  is swept to trace precision/recall/F1 curves. Across configurations, curves cluster tightly around  $\tau = 0$  and around a fixed  $\tau_{MCC}$  (chosen once on a default configuration by maximizing Matthews correlation on a held-out split), with only minor dispersion. Peak F1 consistently occurs near  $\tau \approx 0-0.1$ ; performance degrades only for large positive  $\tau$  where recall collapses (a regime not used operationally). Overall, selection performance is insensitive to reasonable variation in vectorizer, projection, and clustering hyperparameters, supporting  $\tau = 0$  as a default operating point and  $\tau_{MCC}$  as a robustness check.



**Figure 5:** Sensitivity of precision, recall, and F1 to the pruning threshold  $\tau$  across BERTopic hyperparameters (UMAP/HDBSCAN/vectorizer). Vertical lines mark  $\tau = 0$  and a fixed  $\tau_{MCC}$  estimated once on a default configuration. Curves cluster tightly around the operating region ( $\tau \approx 0 - \tau_{MCC}$ ), indicating robustness to hyperparameters.

## 6. Limitations and future direction of work

While the proposed pipeline demonstrates strong performance in identifying thematically coherent and contextually relevant patents, several limitations open avenues for further enhancement. In the expansion and pruning phases, citation networks were leveraged to build a semantically rich candidate set. To avoid redundancy and overfitting, these same features were excluded from the classification phase, opting instead for a single-input BERT architecture that processes concatenated abstracts and titles. This design ensured that distinct sets of information were exploited in different phases, reducing the risk of data leakage or circular logic.

However, treating all textual fields as a single input may dilute domain-specific signals, such as those embedded in CPC codes or reference patterns. A multi-input BERT architecture, where each modality (e.g., abstract, CPC, references) is processed independently before feature fusion, could better preserve the structural and semantic nuances of each information type. For instance, recent work by [13] demonstrates the effectiveness of such architectures in patent retrieval and classification tasks. Future work could therefore explore the development of an unsupervised, multi-input BERT framework tailored to the patent domain to test and capture domain-specific relevance more explicitly.

In the absence of human annotations, a pseudo-labeled set based on overlap with CPC codes is used. While defensible as a proxy for human curation, these labels are not independent of the construction rules; reported metrics should therefore be interpreted as upper-bound estimates. Future work will validate beyond CPC proxies and explore multi-input models that fuse text, classification codes, and citation structure.

## 7. Conclusions

This work introduces an automated, scalable framework for identifying twin (green  $\cap$  digital) patents that integrate green and digital technologies. By combining rule-based seed selection, bidirectional citation expansion, transformer-based embeddings, and topic-guided pruning, the methodology addresses persistent limitations of earlier approaches, including limited coverage and overlap and lack of reproducibility.

Empirical results indicate that the proposed framework yields patent sets with greater topical relevance and improved semantic coherence, as shown by clearer intertopic separation and a KDE shift toward seed topics after pruning. Specifically, the combination of PaECTER embeddings, BERTopic modeling and topic-guided pruning, is shown to enable the effective filtering of heterogeneous patent corpora while minimizing human intervention. Sensitivity analyses indicate small variation across reasonable UMAP/HDBSCAN settings and cosine-similarity thresholds, suggesting a robust pruning score.

Overall, this approach provides a scalable, transparent basis for monitoring innovation dynamics and informing policy in the twin transition.

## 8. Acknowledgments

This study was funded by the European Union - NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP F53C22000760007). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

## 9. Declaration of Generative AI

During the preparation of this work, the author used ChatGPT4.5 in order to perform grammar and spelling check, paraphrase and reword. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

## References

- [1] M. Favot, L. Vesnic, R. Priore, A. Bincoletto, F. Morea, Green patents and green codes: How different methodologies lead to different results, *Resources, Conservation & Recycling Advances* 18 (2023) 200132. doi:10.1016/j.rcradv.2023.200132.
- [2] M. Ghosh, S. Erhardt, M. E. Rose, E. Buunk, D. Harhoff, Paecter: Patent-level representation learning using citation-informed transformers, 2024. doi:10.48550/arXiv.2402.19411.
- [3] M. Grootendorst, BERTopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL: <https://arxiv.org/abs/2203.05794>. arXiv:2203.05794.
- [4] B. Jindra, M. Leusin, The development of digital sustainability technologies by top R&D investors, Technical Report JRC130480, Joint Research Centre (JRC), European Commission, 2022. URL: <https://publications.jrc.ec.europa.eu/repository/handle/JRC130480>. doi:10.2760/150239.
- [5] Z. Griliches, Patent statistics as economic indicators: A survey, *Journal of Economic Literature* 28 (1990) 1661–1707.
- [6] European Patent Office, Guide to the cooperative patent classification (cpc), <https://www.cooperativepatentclassification.org/>, 2016. Accessed 2025-08-25.
- [7] A. Flostrand, L. Pitt, S. Bridson, The delphi technique in forecasting – a 42-year bibliographic analysis (1975–2017), *Technological Forecasting and Social Change* 150 (2020) 119773. doi:10.1016/j.techfore.2019.119773.
- [8] A. Bergeaud, C. Verluise, Identifying technology clusters based on automated patent landscaping, *PLOS ONE* 18 (2023) e0295587. doi:10.1371/journal.pone.0295587.
- [9] R. Capello, C. Lenzi, 4.0 technologies and the rise of new islands of innovation in european regions, *Regional Studies* 55 (2021) 1724–1737. doi:10.1080/00343404.2021.1964698.
- [10] A. Abood, D. Feltenberger, Automated patent landscaping, *Artificial Intelligence and Law* 26 (2018) 103–125. doi:10.1007/s10506-017-9217-1.
- [11] R. Lampe, Strategic citation, *The Review of Economics and Statistics* 94 (2012) 320–333. doi:10.1162/REST\_a\_00146.
- [12] I. Haščič, M. Migotto, Measuring environmental innovation using patent data, OECD Environment Working Papers 89, OECD Publishing, 2015. URL: [https://www.oecd.org/en/publications/measuring-environmental-innovation-using-patent-data\\_5js009kf48xw-en.html](https://www.oecd.org/en/publications/measuring-environmental-innovation-using-patent-data_5js009kf48xw-en.html). doi:10.1787/5js009kf48xw-en.
- [13] 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.
- [14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of NAACL-HLT 2019, Association for Computa-*



- tional Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423/>.
- [15] A. Rietzler, S. Stabinger, P. Opitz, S. Engl, Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4933–4941. URL: <https://aclanthology.org/2020.lrec-1.607/>.
  - [16] J.-S. Lee, J. Hsiang, Patent classification by fine-tuning bert language model, World Patent Information 61 (2020) 101965. URL: <https://www.sciencedirect.com/science/article/pii/S0172219019300742>. doi:<https://doi.org/10.1016/j.wpi.2020.101965>.
  - [17] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, UK, 2008.
  - [18] H. Steck, C. Ekanadham, N. Kallus, Is cosine-similarity of embeddings really about similarity?, 2024. URL: <https://arxiv.org/abs/2403.05440>. doi:10.1145/3589335.3651526. arXiv:2403.05440.
  - [19] L. McInnes, J. Healy, N. Saul, L. Großberger, Umap: Uniform manifold approximation and projection, Journal of Open Source Software 3 (2018) 861. URL: <https://doi.org/10.21105/joss.00861>. doi:10.21105/joss.00861.
  - [20] R. J. G. B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: J. Pei, V. S. Tseng, L. Cao, H. Motoda, G. Xu (Eds.), Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 160–172.
  - [21] R. J. G. B. Campello, D. Moulavi, A. Zimek, J. Sander, Hierarchical density estimates for data clustering, visualization, and outlier detection, ACM Trans. Knowl. Discov. Data 10 (2015). URL: <https://doi.org/10.1145/2733381>. doi:10.1145/2733381.
  - [22] Z. Ye, Y. Geng, J. Chen, J. Chen, X. Xu, S. Zheng, F. Wang, J. Zhang, H. Chen, Zero-shot text classification via reinforced self-training, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 3014–3024. URL: <https://aclanthology.org/2020.acl-main.272/>. doi:10.18653/v1/2020.acl-main.272.
  - [23] D. Zha, C. Li, Multi-label dataless text classification with topic modeling, Knowledge and Information Systems 61 (2019) 137–160. doi:10.1007/s10115-018-1280-0.
  - [24] A. Ratner, S. H. Bach, P. Varma, C. Ré, et al., Snorkel: rapid training data creation with weak supervision, The VLDB Journal 29 (2020) 709–730. doi:10.1007/s00778-019-00552-1.
  - [25] D. Sculley, Web-scale k-means clustering, in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 1177–1178. URL: <https://doi.org/10.1145/1772690.1772862>. doi:10.1145/1772690.1772862.
  - [26] S. P. Lloyd, Least squares quantization in pcm, IEEE Trans. Inf. Theory 28 (1982) 129–136.
  - [27] C. Sievert, K. Shirley, LDAvis: A method for visualizing and interpreting topics, in: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 63–70. doi:10.3115/v1/W14-3110.
  - [28] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: <https://aclanthology.org/D19-1410/>.
  - [29] H. Gonen, Y. Goldberg, Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 609–614. URL: <https://aclanthology.org/N19-1061/>.
  - [30] G. Salton, A. Wong, C. S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (1975) 613–620. URL: <https://doi.org/10.1145/361219.361220>. doi:10.1145/361219.361220.
  - [31] European Patent Office, Patstat autumn 2024: Worldwide patent statistical database, <https://www>.

epo.org/en/searching-for-patents/business/patstat, 2024. Accessed 2025-08-25.

- [32] European Patent Office, PATSTAT 2024 Autumn Edition Data Catalog, v5.24, Technical Report, European Patent Office, 2024. URL: <https://link.epo.org/web/searching-for-patents/business/patstat/data-catalog-patsat-global-en.pdf>, accessed 2025-01-10.