CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21880

Longitudinal Analysis of Patents in Brazil between 2000 and 2025 Using Machine Learning and Natural Language Processing

Priscila Rezende da Costa[1], Camila Naves Arantes[2], Carlos Eduardo Costa de Glavina Reino[3]

[1]PPGA ESPM, ICT ESPM, Brazil, [2]INTELI, Brazil, [3]UNESP, Brazil.

How to cite: da Costa, P. R.; Naves Arantes, C.; de Glavina Reino, C. E. C. (2026). Longitudinal Analysis of Patents in Brazil between 2000 and 2025 Using Machine Learning and Natural Language Processing. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 26-28 June 2026. https://doi.org/10.4995/CARMA.2026.21880

Abstract

This study analyses the technological dynamics of patents published in the Brazilian jurisdiction between 2000 and 2025, integrating descriptive statistics, machine learning, and natural language processing. The dataset comprises 113,861 records extracted from Lens.org, including bibliographic, technological, and legal information. Indicators of activity, technological concentration, and impact were estimated, alongside semantic space construction using TF IDF and topic modelling through latent Dirichlet allocation. Supervised and unsupervised techniques were applied to estimate grant probability, decision time, and technological clustering. Results indicate expansion concentrated in an intermediate period, subsequent contraction, low technological concentration, and dissociation between patent volume and citation impact. Intermediate clusters display higher semantic centrality and are associated with digital and energy related domains. Findings suggest that the evolution of the Brazilian patent system occurs through selective recombination across technological domains, with implications for innovation policy and intellectual property portfolio management.

Keywords: *Patents; Machine learning; Natural language processing; Technological convergence; Innovation indicators.*

1. Introduction

Patent databases are used to measure technological activity, analyze innovation trajectories, and examine knowledge diffusion through citations and classifications such as CPC and IPC (Griliches, 1990; Hall, Jaffe, & Trajtenberg, 2001). The scale and textual structure of patent documents require computational methods for large-scale extraction. Reviews describe the use of natural language processing in patent analysis, including classification, semantic similarity, information retrieval, and topic modeling based on transformer and embedding models, and

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

identify gaps in methodological integration and empirical validation across institutional contexts (Jiang et al., 2025; Lin & Chou, 2025).

Studies report performance gains in classification using domain-adapted BERT models at the CPC and IPC levels (Lee & Hsiang, 2020), deep learning approaches combining text and metadata (Sofean, 2021), semantic embeddings such as Sentence BERT for similarity and classification (Bekamiri, Hain, & Jurowetzki, 2024), and technology mapping through text mining and network analysis (Wang et al., 2023). There remains a gap in integrating traditional patent indicators with machine learning and natural language processing in a longitudinal analysis of patents in the Brazilian jurisdiction (Lin & Chou, 2025; OECD, 2025).

The research question examines how the technological dynamics of patents published in the Brazilian jurisdiction evolve between 2000 and 2025 and how machine learning and natural language processing identify technological and structural patterns. The study integrates descriptive statistics, supervised and unsupervised models, topic modeling, and similarity measures to analyze temporal evolution, technological distribution, concentration, decision cycle, legal status, and technological clusters. It connects the literature on patent indicators with longitudinal evidence from Brazil and applies AI methods described in recent reviews (Jiang et al., 2025; Lin & Chou, 2025), providing a reproducible procedure for technological monitoring to support intellectual property management and innovation policy.

2. Method

3. Results

The study adopted a quantitative, descriptive, and longitudinal design based on secondary data from patents published in the Brazilian jurisdiction between 2000 and 2025. The database includes bibliographic, technological, and legal information, such as dates, classifications, holders, inventors, citations, patent family size, and legal status, extracted from Lens.org. Data were standardized, verified, deduplicated, and treated for missing values. Temporal variables were organized by filing and publication year, and technological classifications were obtained from CPC and IPCR codes at the subclass level. The use of patent data as a proxy for technological activity follows the literature on technological indicators (Griliches, 1990; Hall, Jaffe, & Trajtenberg, 2001).

Indicators comprised annual filings and publications, technological distribution, citations, patent family size, and legal status. Technological intensity and concentration were measured using CPC and IPCR frequencies and the Herfindahl Hirschman index. Supervised models estimated grant probability and time to decision using decision trees and ensemble methods with cross validation (Breiman, 2001), while unsupervised clustering identified technological

Article in Press

groups. Text analysis applied term frequency and inverse document frequency weighting (Salton & Buckley, 1988) and latent Dirichlet allocation (Blei, Ng, & Jordan, 2003). Model performance was evaluated using accuracy, precision, recall, mean squared error, and area under the ROC curve (Hastie, Tibshirani, & Friedman, 2009). Procedures were conducted in a documented computational environment to ensure reproducibility.

3.1. Semantic Space and Properties of Vector Representation

The semantic space was constructed from 113,861 patents published in Brazil between 2006 and 2024, using concatenated titles and abstracts. After capping text length at the 99th percentile of 320 tokens, the corpus showed a minimum of 1 token, first quartile of 63, median of 127, mean of 123.06, and third quartile of 178 tokens, indicating informational heterogeneity consistent with technological recombination (Fleming, 2001) and with the use of patent text and citations as indicators of knowledge flows (Hall, Jaffe, & Trajtenberg, 2001). Vector representation was estimated using TF IDF weighting with unigrams and bigrams, a maximum of 80,000 attributes, and a minimum frequency of 10 documents. In a random sample of 15,000 patents, the space reached d distinct attributes, reflecting high dimensionality and sparsity, in line with evidence on dispersed patent vector spaces under term frequency models (Arts, Cassiman, & Gomez, 2018; Lee & Hsiang, 2020). This structure supports clustering and topic modeling and is consistent with studies showing that large scale vector representations enable identification of semantic proximity, technological trajectories, and related diversification (Balland, Boschma, Crespo, & Rigby, 2019; Rigby, Roesler, Kogler, & Boschma, 2022), while dense vectors and transformer models improve semantic similarity and thematic classification in large patent databases (Lee & Hsiang, 2020; Bekamiri, Hain, & Jurowetzki, 2024).

3.2. Identification and Characterization of Technological Clusters

The application of unsupervised clustering to the TF IDF representation of 113,861 patents generated 10 technological clusters and revealed asymmetries in size, centrality, and semantic position, combining semantic network metrics and cluster topology in line with approaches based on textual embeddings (Bekamiri, Hain, and Jurowetzki, 2024; Lee and Hsiang, 2020). The largest clusters account for about 40 percent of patents, show low betweenness, and concentrate CPC subclasses in chemistry, mechanics, and engineering, consistent with evidence on thematic specialization and mature trajectories with limited cross domain integration (Kogler, Rigby, and Tucker, 2013; Arts, Cassiman, and Gomez, 2018; Rigby, Roesler, Kogler, and Boschma, 2022). Intermediate clusters, representing 10 to 15 percent of patents, present higher betweenness and degree and are dominated by CPC codes such as H04W, H04L, and G06N, linking communication networks, information systems, and automation, in line with studies on cross sectoral convergence and similarity networks (Yoon and Kim, 2011; Blei, Ng, and Jordan, 2003; Balland et al., 2019; Bekamiri et al., 2024; Ayaou et al., 2025).

Article in Press

Peripheral clusters account for a smaller share, show low connectivity and limited CPC diversity, and may correspond to emerging trajectories that can gain relevance through recombination (Boschma, 2017; Balland, Boschma, and Frenken, 2015; Rigby et al., 2022; Arts et al., 2018). Cluster relevance does not correlate directly with patent volume, indicating that semantic integration and topological position provide more information than size alone (Kogler et al., 2013; Bekamiri et al., 2024; Arts et al., 2018; Lee and Hsiang, 2020), and the predominance of international patent families indicates integration with global innovation flows and the role of international networks in technological diffusion and specialization (de Rassenfosse et al., 2013; Criscuolo and Verspagen, 2008).

3.3. Explanatory Capacity of Models for Technological Impact Dimensions

The explanatory models applied to the three dimensions of technological impact, forward citations, family size, and geographic reach, show different levels of explanatory capacity. The model for forward citations, which measures how often a patent is cited, presented an R^2 of approximately 0.118 and an RMSE of 7.824, indicating limited but statistically significant explanatory power. This result is consistent with studies that describe citations as noisy measures, sensitive to temporal, institutional, and knowledge diffusion factors, particularly in dynamic fields (Griliches, 1990; Hall, Jaffe & Trajtenberg, 2001), and with evidence on the complexity of citation metrics in patent research (Kogan, Papanikolaou, Seru & Stoffman, 2017) (Table 1).

The model for family size presented an R^2 of approximately 0.541 and an RMSE of 10.702, indicating moderate explanatory power regarding the international extension of protection. The literature interprets patent family size as an indicator of expected value and appropriation strategy under technological uncertainty and global competition (Teece, 1986; Harhoff, Scherer & Vopel, 2003). Empirical studies show that family size is a consistent predictor of technological value and expected economic returns when combined with market and institutional variables (de Rassenfosse, Dernis, Guellec, Picci & van Pottelsberghe de la Potterie, 2013; Criscuolo & Verspagen, 2008) (Table 1).

The model for geographic reach, measured by the number of jurisdictions in the patent family, presented an R^2 of approximately 0.322 and an RMSE of 3.897, indicating a moderate relationship between the explanatory variables and the territorial dispersion of protection. Although structural patent characteristics are associated with geographic extent, part of the variation is related to corporate strategies, sectoral conditions, and market factors. Research on patent internationalization argues that geographic reach reflects strategic decisions about market entry and protection against foreign competitors (de Rassenfosse et al., 2013; Criscuolo & Verspagen, 2008), and recent studies indicate that region selection depends on industry characteristics and regulatory regimes (Kogan et al., 2017) (Table 1). Overall, family size and geographic reach show statistically relevant explanatory power and are consistent with

Article in Press

theoretical expectations, while forward citations require more complex specifications, including nonlinear models or network variables, to better capture their behavior (Arts, Cassiman & Gomez, 2018; Bekamiri, Hain & Jurowetzki, 2024). These findings indicate the need for multiple analytical approaches to assess technological impact using patent data (Table 1).

Table 1. Summary table of the performance of the explanatory models. Source: Research data.

Impact dimension	Target variable	Approx. R^2	Approx. RMSE
Forward citations	Cited by patent count	0.118	7.824
Family size	Simple family size	0.541	10.702
Geographic scope	Number of family jurisdictions	0.322	3.897

3.4. Technological Categories and Macrotechnologies in the Brazilian Patent Database Between 2006 and 2024

The decomposition of the Brazilian database of 113,861 patents, based on the co-occurrence of CPC subclasses, relative frequency analysis, and semantic similarity between titles and abstracts, indicates a technological structure marked by sectoral dispersion and interdependence. The Herfindahl-Hirschman index for the main subclasses was 0.0049, indicating low concentration and related variety, consistent with studies linking diversified innovation systems to incremental recombination between related domains (Kogler, Rigby, & Tucker, 2013; Balland, Boschma, & Frenken, 2015) (Table 2). CPC sections G and H, related to physics, computing, communication, and electrical technologies, as well as section Y, related to climate mitigation, show higher participation and relative growth over time despite fluctuations in patent volume. The co-occurrence between electrical and mechanical subclasses indicates integration between physical and digital systems, associated with digitization and electrification processes (Fouquet & Hippe, 2022; Jiang et al., 2025) (Table 2).

Subclasses G06, H04, and G05 present higher centrality in the semantic similarity network, measured by degree and betweenness, connecting distinct technological groups and performing an articulation function. Evidence based on textual embeddings shows that digital technologies tend to occupy intermediate network positions, facilitating convergence (Bekamiri, Hain, & Jurowetzki, 2024; Lee & Hsiang, 2020). Domains with higher centrality also show greater probability of future diversification (Rigby, Roesler, Kogler, & Boschma, 2022) (Table 2). In contrast, materials, manufacturing, and mechanical engineering show high patent volume but lower betweenness centrality, indicating specialization with limited bridging across domains. This distinction between volume and structural position aligns with research showing that technological relevance depends on relational position rather than inventive intensity alone (Hall, Jaffe, & Trajtenberg, 2001; Arts, Cassiman, & Gomez, 2018) (Table 2).

Article in Press

Longitudinal analysis identifies three dynamics: growth of digital and energy technologies, stability of traditional industrial domains, and increasing semantic interdependence between physical and digital areas. These findings align with recombination models in which new trajectories emerge from combining knowledge in related regions (Fleming, 2001; Kogler et al., 2013) and with studies showing that technological convergence occurs through cumulative and selective mechanisms rather than substitution (Balland et al., 2019; Huang et al., 2025) (Table 2). Overall, the Brazilian technological structure reflects the coexistence of consolidated domains with high internal density and transversal domains with high connectivity. Convergence occurs through recombination between digital, energy, and industrial areas, indicating restructuring of the national technological system (Table 2).

Table 2. Macrotechnologies, relative share, and structural position. Source: Research data.

Macrotechnology	Relative share in the database	Centrality in the semantic network	Temporal dynamics	Structural function
Electrification and energy	Relative growth	Medium	Integration expansion	Physical-electrical articulation
Digital technologies and control	Relative growth	High	Structural consolidation	Interdomain bridge
Materials and manufacturing	Stable	Low	Incremental	Productive specialization
Traditional mechanical engineering	Stable to slightly decreasing	Low	Consolidation	Consolidated domain
Systemic integration	Moderate	High	Recombinatory	Technological convergence

4. Conclusion

The technological dynamics observed in the Brazilian patent database between 2006 and 2024 are characterized not by homogeneous cumulative expansion, but by a structural reconfiguration of the technological space. The analysis shows that the evolution of the patent system occurs through selective recombinations among digital, energy, and industrial domains, organized in semantic groupings with distinct intermediation functions. Domains related to digital technologies and electrification occupy central positions in the similarity network, while traditional industrial areas retain significant filing volume but perform a less central structural role. The results indicate that technological relevance does not necessarily follow the growth of inventive effort, as the dissociation among filings, citations, and geographic scope reveals limitations in linear approaches based solely on patent intensity.

The practical implications shift decision-making from a quantitative to a relational perspective. For firms, technological advantage depends on operating at intersections between domains, particularly between digital technologies and physical systems, rather than on the isolated

Article in Press

accumulation of technological assets. For policymakers, sector-based incentive instruments may capture only part of innovation dynamics, whereas prioritizing technologies with integrating functions and recombination potential is more consistent with observed patterns. For regulators, the interaction between institutional elements, technical requirements, and technological trajectories supports the need for adaptive regulatory approaches aligned with structural changes.

The study has empirical and methodological limitations. Using patents as the unit of analysis assumes that innovation is observable through formal appropriation mechanisms, which may underrepresent tacit knowledge, organizational processes, and non-patented innovations. Impact metrics such as citations and family size reflect strategic and institutional contexts, limiting their capacity to measure technological value under uncertainty. Methodologically, the construction of semantic space and clustering parameters influences the identified structure, so alternative parameter settings may generate different but plausible configurations, requiring interpretive caution and transparency.

Future research may treat technological convergence as a dynamic and contingent process. Longitudinal analyses using moving windows can examine how intermediation functions emerge, consolidate, or decline. Integrating patent data with information on investments, certifications, market performance, and collaborative networks can clarify how recombinatorial trajectories develop into productive systems. Methodological advances combining multiple semantic representations, explainability techniques in machine learning models, and cross-validation procedures can reduce analytical dependence and increase robustness. These developments contribute to understanding the Brazilian technological system as a process of structural reorganization of intersectoral innovation.

References

Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal, 39*(1), 62–84. https://doi.org/10.1002/smj.2699

Ayaou, I., et al. (2025). PatenTEB: A patent text embeddings benchmark for technological landscaping. *arXiv*. https://arxiv.org/pdf/2510.22264

Balland, P. A., Boschma, R., & Frenken, K. (2015). Proximity and innovation: From statics to dynamics. *Regional Studies, 49*(6), 907–920. https://doi.org/10.1080/00343404.2014.883598

Balland, P. A., Boschma, R., Crespo, J., & Rigby, D. L. (2019). Smart specialization policy in the European Union: Relatedness, knowledge complexity and regional diversification. *Regional Studies, 53*(9), 1252–1268. https://doi.org/10.1080/00343404.2018.1437900

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

Article in Press

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Boschma, R. (2017). Relatedness as driver of regional diversification: A research agenda. *Regional Studies, 51*(3), 351–364. https://doi.org/10.1080/00343404.2016.1254767

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breschi, S., Lissoni, F., & Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy, 32*(1), 69–87. https://doi.org/10.1016/S0048-7333(02)00004-5

Criscuolo, P., & Verspagen, B. (2008). Does it matter where patent citations come from? *Research Policy, 37*(10), 1892–1908. https://doi.org/10.1016/j.respol.2008.07.007

de Rassenfosse, G., Dernis, H., Guellec, D., Picci, L., & van Pottelsberghe de la Potterie, B. (2013). The worldwide count of priority patents: A new indicator of inventive activity. *Research Policy, 42*(3), 720–737. https://doi.org/10.1016/j.respol.2012.11.002

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117–132. https://doi.org/10.1287/mnsc.47.1.117.10671

Fouquet, R., & Hippe, R. (2022). Twin transitions of decarbonisation and digitalisation. *Energy Research & Social Science, 91*, 102736. https://doi.org/10.1016/j.erss.2022.102736

Geels, F. W. (2002). Technological transitions as evolutionary reconfiguration processes. *Research Policy, 31*(8–9), 1257–1274. https://doi.org/10.1016/S0048-7333(02)00062-8

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature, 28*(4), 1661–1707.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools. *NBER Working Paper No. 8498*.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics, 36*(1), 16–38.

Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy, 32*(8), 1343–1363. https://doi.org/10.1016/S0048-7333(02)00124-5

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Huang, Y., Xu, S., Cai, S., & Lü, L. (2025). Uncovering multi technology convergence patterns with hypergraphs. *arXiv preprint*.

Jiang, L., et al. (2025). Natural language processing in the patent domain: A survey. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-025-11168-z

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics, 132*(2), 665–712. https://doi.org/10.1093/qje/qjw040

Kogler, D. F., Rigby, D. L., & Tucker, I. (2013). Mapping knowledge space and technological relatedness. *Economic Geography, 89*(4), 373–392. https://doi.org/10.1111/ecge.12018

Lee, J. S., & Hsiang, J. (2020). Patent classification by fine tuning BERT language model. *World Patent Information, 61*, 101965. https://doi.org/10.1016/j.wpi.2019.101965

Article in Press

Lin, T.-Y., & Chou, L.-C. (2025). A systematic review of artificial intelligence applications and methodological advances in patent analysis. *World Patent Information, 82*, 102383. https://doi.org/10.1016/j.wpi.2025.102383

OECD. (2025). *Identifying emerging AI technologies using patent data: A semi-automated approach* (Technical paper). OECD Publishing.

Rigby, D. L., Roesler, C., Kogler, D. F., & Boschma, R. (2022). Do EU regions benefit from smart specialization principles. *Regional Studies, 56*(12), 2059–2072. https://doi.org/10.1080/00343404.2022.2032628

Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sofean, M. (2021). Deep learning based pipeline with multichannel inputs for patent classification. *World Patent Information, 66*, 102060. https://doi.org/10.1016/j.wpi.2021.102060

Teece, D. J. (1986). Profiting from technological innovation. *Research Policy, 15*(6), 285–305. https://doi.org/10.1016/0048-7333(86)90027-2

Wang, J., et al. (2023). Tracking and predicting technological knowledge interactions between artificial intelligence and wind power: Multimethod patent analysis. *Advanced Engineering Informatics, 58*, 102177. https://doi.org/10.1016/j.aei.2023.102177

Yoon, B., & Kim, S. (2011). Identifying technological competition trends for R&D planning using dynamic patent maps. *Technological Forecasting and Social Change, 78*(3), 477–492. https://doi.org/10.1016/j.techfore.2010.08.003

Article in Press