Lightweight Adherence Prediction and Justification for Government Strategic Priorities

Matheus Utino[1], **Marcos Gôlo**[1], **Marcelo Turine**[1,2], **Ricardo Marcacini**[1]

[1]Institute of Mathematical and Computer Sciences, University of São Paulo

[2]Faculty of Computer Science, Federal University of Mato Grosso do Sul

matheusutino@usp.br, marcosgolo@usp.br

marcelo.turine@ufms.br, ricardo.marcacini@icmc.usp.br

Abstract. *The Brazilian National Postgraduate System plays a strategic role in scientific development and policymaking. The CAPES National Agenda organizes 470 strategic themes into 20 macro-themes, reflecting priorities from Brazil's 27 Federative Units. Given the large annual volume of academic outputs, identifying their alignment with these priorities is essential. This study evaluates lightweight machine learning models as scalable alternatives to LLMs for adherence prediction and justification. Using 42,026 academic outputs and nine classifiers, Bag-of-Words with Random Forest achieved competitive performance (F_1 of 0.755), outperforming semantic embeddings. A fine-tuned PTT5-base produced explanations highly similar to human references.*

1. Introduction

The formation of postgraduate students (master's and doctoral levels) within the Brazilian National Postgraduate System (SNPG) constitutes a structural pillar for the planning, implementation, and evaluation of public policies in Brazil. The SNPG plays a fundamental role in scientific and technological development by preparing highly qualified human resources for research and innovation. To ensure the effectiveness of postgraduate education in addressing societal challenges and promoting national development, it must not only foster knowledge production but also align with labor market demands, employability needs, and the social, economic, human, and environmental challenges faced across Brazilian territories [CAPES 2021, CAPES 2023].

In this context, the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*, Ministry of Education (CAPES/MEC) has recently introduced the *Agenda Nacional para a Formação de Pessoal de Nível Superior*[1], which establishes a strategic framework by organizing 470 declared strategic themes across all 27 Brazilian states to guide public investment and regional development [CAPES 2025]. These themes are structured into 20 macro-themes that guide the induction, funding, and evaluation of postgraduate education in Brazil, such as, M1 – Sustainable and Regional Development; M2 – Public Policies, Management and Governance; M4 – Digital Transformation, Information and Communication Technologies, and Artificial Intelligence; M9 – Health, Quality of Life and Technologies; and M10 – Biodiversity, Biotechnology and Bioeconomy.

[1]https://agendanacional.capes.gov.br/

Within this framework, SNPG data for 2025[2] indicate that Brazil has 457 science and technology institutions (ICTs) offering 4,635 postgraduate programs, involving 111,699 faculty advisors and a total of 432,888 students (master's and doctoral levels), as well as 1,164,454 intellectual outputs [CAPES 2026]. Against this backdrop, identifying the volume and profile of academic output aligned with these priorities becomes critical for public managers responsible for research funding, academic partnerships, and science-policy integration [Jin and Mihalcea 2022, Jiang et al. 2023]. Figure 1 illustrates the task of predicting adherence between researchers and strategic themes.

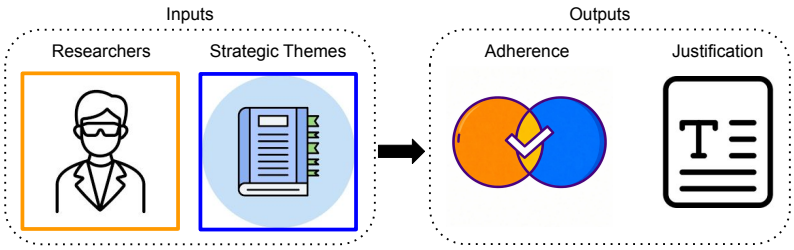


Figure 1. Illustration of the prediction of researcher and topic adherence.

Given the scale of postgraduate defended works, manual assessment of alignment with the national agenda is unscalable, making automated classification a prerequisite for evidence-based policymaking [ICMC 2025, Landhuis 2016, Barbier and Jeangirard 2025, Brasil 2021]. Large Language Models (LLMs) have emerged as the state of the art across a wide range of natural language processing tasks [Annepaka and Pakray 2025], and represent a promising alternative for this task, as they perform semantic mapping between academic productions and strategic themes while simultaneously producing natural language justifications that support transparency and auditability. Nevertheless, their direct deployment at scale incurs substantial computational costs, rendering them infeasible for resource-constrained public-sector environments [Sakiyama et al. 2025]. Prior work has largely framed adherence as a binary filtering task, focusing exclusively on the classification stage, whether through embedding-based classifiers [Utino and Gôlo 2025, Miranda et al. 2025] or one-class learning [Gôlo and Utino 2025], without addressing the generation of formal textual justifications required by public governance regulations.

To address these limitations, this paper introduces a two-stage pipeline designed for scalable public-sector deployment. In the first stage, adherence labels generated by a reference LLM are used to train lightweight classifiers over text representations, enabling accurate adherence prediction at less CPU inference cost. In the second stage, a fine-tuned sequence-to-sequence model conditioned on the predicted label generates explicit natural language justifications, fulfilling the auditability requirements mandated by public governance regulations. The contributions of this work are:

- **Lightweight Adherence Classification:** Sparse representations (BoW) with Random Forest effectively classify academic productions against strategic themes, outperforming costly dense embeddings.

[2]https://sucupira-v2.capes.gov.br/painel

- **Auditable Justification Generation:** A dedicated PTT5-based generative module synthesizes natural language justifications conditioned on predicted labels, achieving high semantic fidelity (0.90) without relying on massive LLMs.
- **Open Source and Reproducibility:** The complete pipeline, experimental logs, and prompt templates are publicly available[3].

2. Problem Definition

Let the raw data consist of instances (x_i, t_i, a_i, e_i), where $x_i \in \mathcal{X}$ represents an academic production and $t_i \in \mathcal{T}$ represents a strategic public policy theme defined by a Federative Unit. Each pair is originally associated with a qualitative adherence assessment $a_i \in \{\text{Non-Adherent}, \text{Adherent}\}$ and a corresponding natural language justification $e_i \in \mathcal{E}$, where $\mathcal{E}$ denotes the space of textual sequences. Consequently, we define our formal working dataset as $\mathcal{D} = \{(x_i, t_i, a_i, e_i)\}_{i=1}^{N}$, where N denotes the total number of instances. Prior to classification, each raw input pair (x_i, t_i) is mapped to a d-dimensional feature vector through a representation function $\phi : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^d$, such that $v_i = \phi(x_i, t_i)$. The proposed framework addresses two interdependent tasks based on $\mathcal{D}$:

1. **Adherence Classification:** Learning a predictive function $f : \mathbb{R}^d \to \{\text{Non-Adherent}, \text{Adherent}\}$ that maps the feature representation to an adherence label $\hat{a}_i = f(\mathbf{v}_i)$.
2. **Justification Generation:** Learning a conditional text generation model $g : \mathcal{X} \times \mathcal{T} \times \{\text{Non-Adherent}, \text{Adherent}\} \to \mathcal{E}$ that produces a natural language explanation $\hat{e}_i = g(x_i, t_i, \hat{a}_i)$ articulating the reasoning behind the predicted adherence level.

3. Related Work

Recent studies have addressed the classification of academic profiles and strategic public themes by exploring diverse embedding representations and similarity aggregation techniques. For instance, researcher-topic affinity has been formulated as a binary classification task, demonstrating that pairing dense embeddings with lightweight classifiers optimizes the identification of highly relevant pairs [Utino and Gôlo 2025]. Similarly, fine-grained affinity estimation strategies have been proposed to compute and aggregate embedding similarities across distinct textual fields, utilizing specific weighting mechanisms to generate a final adherence score [Miranda et al. 2025].

To tackle the inherent complexities of mapping academic datasets, such as severe class imbalance, LLMs have been employed with a Balanced Loss function during fine-tuning [Sakiyama et al. 2025]. While LLMs offer strong predictive performance and the ability to generate natural-language justifications, their deployment incurs substantial computational costs and high carbon emissions, making them infeasible for resource-constrained environments. As an alternative to heavy models and exhaustive negative sampling, a one-class learning paradigm via Support Vector Data Description (SVDD) has been applied to model the region of interest [Gôlo and Utino 2025]. Although this approach has the advantage of training solely on positive instances and provides visual interpretation, it lacks the capacity to generate the formal textual justifications required by public governance regulations.

[3]https://github.com/Matheusutino/LASDiGov-2026-lightweight-gov-adherence

Our work distinguishes itself by proposing a two-stage pipeline that relies entirely on lightweight methods for both classification and justification generation. First, we demonstrate that highly accurate adherence prediction can be achieved with lightweight representations and computationally inexpensive classifiers, thereby circumventing the need for large models. Second, rather than relying on heavy LLMs for explainability or settling for mere visual interpretations, we introduce a dedicated, lightweight generative module that synthesizes explicit natural-language justifications for each predicted label. This approach ensures that the system scales efficiently in resource-constrained public-sector environments while providing the formal, text-based reasoning required for transparent, evidence-based policymaking.

4. Methodology and Experimental Setup

This section describes the experimental design adopted to validate the proposed pipeline, encompassing the dataset construction, text representation strategies, classification algorithms, the justification generation module, and the evaluation protocol.

4.1. Dataset

This study relies on a publicly available dataset prepared for the LeanDL-HPC 2025 Challenge [Rigo and Assunção 2025], comprising $N = 42{,}026$ records of theses and dissertations defended in Brazil in 2023. Each record contains rich metadata about the academic production—including title, research line, area of concentration, project name, and abstracts in both Portuguese and English—paired with a strategic theme, its qualitative assessment, and a natural language justification. The dataset spans 438 unique strategic themes across different Federative Units, reflecting the heterogeneity of regional governmental agendas. To construct the input for the models, the raw textual data for the pair (x_i, t_i) is consolidated by concatenating the strategic theme, its associated keywords, the production title, the project name, and both the Portuguese and English abstracts. This design ensures that the representation function ϕ has access to both the specific governmental context and the comprehensive semantic content of the academic production.

The dataset $\mathcal{D}$ is partitioned into a training set (80%) and a held-out test set (20%) via stratified sampling, preserving the original class distribution of the binary label a_i. To evaluate the models' sensitivity to data availability and assess performance variations, the training set is further subdivided into progressively smaller fractions. The test set remains fixed across all experiments to provide a consistent baseline for evaluation.

4.2. Text Representation Strategies

Three distinct representation strategies are evaluated to instantiate the extraction function ϕ, mapping the input pair (x_i, t_i) into a continuous feature vector $\mathbf{v}_i \in \mathbb{R}^d$:

- **Bag-of-Words (BoW):** Represents documents as term-frequency vectors over a fixed vocabulary, fitted on the training split to avoid leakage. Preprocessing includes lowercase normalization, Portuguese stopword removal, and a limit of $d = 1{,}000$ features.
- **TF-IDF:** Weights terms by inverse document frequency, reducing the impact of frequent, less informative terms. The same preprocessing and vocabulary constraints as BoW are applied.

- **Semantic Embeddings:** Encodes texts into dense vectors capturing semantic relations between academic content and governmental themes. We use the ibm-granite/granite-embedding-278m-multilingual[4] [Awasthy et al. 2025], which supports Portuguese–English inputs.

4.3. Classification Algorithms

To learn the predictive function $f : \mathbb{R}^d \rightarrow \{\text{Non-Adherent}, \text{Adherent}\}$ that maps the feature vector $\mathbf{v}_i$ to the adherence label $\hat{a}_i$, nine classifiers are evaluated across all three representation strategies. The selected models cover a broad methodological spectrum: **Logistic Regression** and **Ridge Classifier** as linear probabilistic and regularized discriminative models; **SGD Classifier** (hinge loss) and **Linear SVC** as margin-based methods; **Passive-Aggressive Classifier** as an online learning approach; **Nearest Centroid** as a distance-based baseline; and **Decision Tree**, **Random Forest**, and **Histogram-based Gradient Boosting** representing the tree-based family, ranging from a single tree to strong ensemble methods. This selection enables a systematic comparison across model families with substantially different inductive biases, computational profiles, and interpretability characteristics, all of which are relevant considerations for government deployment.

4.4. Justification Generation

The justification generation component is formulated as a conditional sequence-to-sequence task that learns the generative function $g : \mathcal{X} \times \mathcal{T} \times \{\text{Non-Adherent}, \text{Adherent}\} \rightarrow \mathcal{E}$. Its primary objective is to synthesize a coherent natural language explanation, $\hat{e}_i$, that articulates the reasoning behind the predicted adherence level for a given academic production and strategic theme. Thus, we employ the PTT5-base model, a variant of the T5 architecture specifically pre-trained on a large-scale Brazilian Portuguese corpus[5][Carmo et al. 2020].

The generative process is guided by a highly structured input prompt, p_i, which maps the tuple $(x_i, t_i, \hat{a}_i)$ into a continuous textual sequence. Specifically, the prompt p_i concatenates the classification outcome (the binary label $\hat{a}_i$ mapped to its natural language equivalent), the governmental context (the strategic theme t_i and its official keywords), and the comprehensive academic profile (x_i, which encompasses the production title, project name, research keywords, and abstracts). During the fine-tuning phase, the prompt is conditioned on the ground-truth label a_i, allowing the model to learn the underlying mapping between the text features and the reference justification e_i. At inference time, this label is replaced by the prediction $\hat{a}_i$ generated by the classification module, producing the final explanation $\hat{e}_i = g(x_i, t_i, \hat{a}_i)$. This conditioning mechanism ensures that the synthesized text remains strictly faithful to the automated decision, providing an auditable rationale tailored to the system's output.

4.5. Evaluation Metrics

First, the adherence classification module is assessed using the F_1-macro to robustly account for class imbalances. Second, the generative quality of the PTT5 model is evaluated against human references using lexical overlap metrics (ROUGE and SacreBLEU), complemented by a semantic similarity score computed as the cosine similarity between the

[4]https://huggingface.co/ibm-granite/granite-embedding-278m-multilingual
[5]https://huggingface.co/unicamp-dl/ptt5-base-portuguese-vocab

embeddings of the generated and reference justifications, encoded by the Granite model. A qualitative analysis of a representative sample of generated justifications is also conducted to assess practical coherence and readability beyond what automatic metrics can capture. Finally, to address the infrastructure constraints typical of the public sector, the computational cost of the classification stage is evaluated by comparing training and inference wall-clock times across all combinations of algorithms and representations.

5. Results and Discussion

To understand how the volume of available training data affects the predictive capability and stability of the classifiers, we analyzed the learning curves for the F_1-macro. Figure 2 illustrates the performance trajectories of the nine evaluated algorithms trained on incremental subsets ranging from 10% to 100% of the training data (i.e., from 10% to 100% of the 80% training split, with the remaining 20% held out for evaluation).

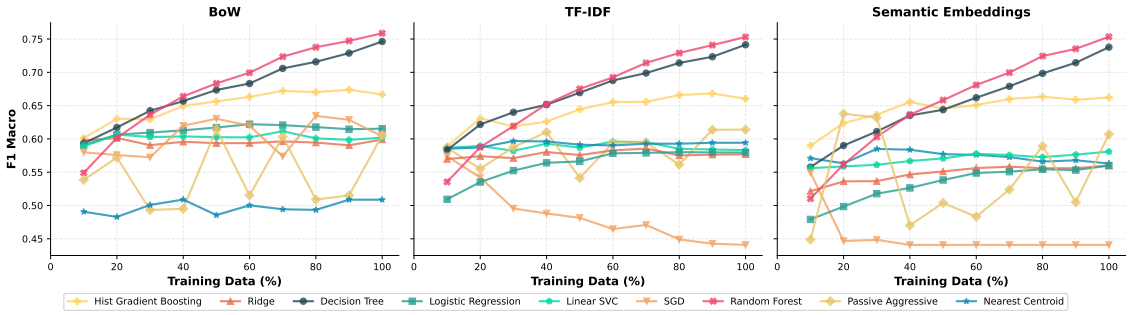


Figure 2. Learning curves depicting F_1-macro across different training data proportions (10% to 100%) for BoW, TF-IDF, and embeddings representations.

Analyzing trajectories as the dataset size increases reveals distinct learning behaviors across algorithmic families. Tree-based ensemble methods, specifically the Random Forest and Decision Tree classifiers, exhibit a strong, near-linear increase in performance as more training instances are added. They show continuous performance gains across the incremental splits, with no clear signs of early plateauing, suggesting a high capacity to leverage additional data to map the dataset's underlying patterns. On the other hand, the Histogram-based Gradient Boosting algorithm demonstrates remarkable early data efficiency. It achieves near-peak predictive performance even when restricted to 10% to 20% of the training data. However, its learning curve flattens rapidly thereafter, indicating that further data injection yields marginal returns for this specific architecture.

In contrast to the tree-based approaches, linear and margin-based classifiers, such as Logistic Regression, Ridge, and Linear SVC, present relatively flat learning curves from the outset. These models stabilize quickly, suggesting that their inherent representational capacity is reached early in the training continuum and that adding new instances does not yield meaningful performance improvements. Furthermore, the incremental data injection exposes significant behavioral instability in certain approaches. The Passive-Aggressive classifier exhibits severe volatility across all sequential data splits. More notably, the SGD classifier displays an anomalous degradation trajectory; its predictive performance systematically declines as the volume of training data increases. This phenomenon suggests that the SGD formulation becomes increasingly susceptible to optimization challenges or overfitting as the empirical distribution grows in size/complexity.

These empirical observations carry significant practical implications for the deployment of artificial intelligence in public administration. The sustained performance gains exhibited by tree-based ensembles suggest they are well-suited to long-term, large-scale government systems, such as the CAPES repository, where the volume of academic data continues to expand. As these national databases grow, such models will organically improve their ability to map strategic themes, maximizing the return on investment in data infrastructure. On the other hand, the remarkable early data efficiency demonstrated by Gradient Boosting architectures offers a strategic advantage for state agencies or municipal governments operating in data-scarce environments. For newly formulated public policies or emerging strategic themes with limited historical records, such algorithms can provide reliable decision support without requiring massive initial annotation efforts. Finally, the instability and degradation observed in models such as SGD and Passive-Aggressive highlight a critical risk for public-sector applications: a lack of predictive consistency can compromise administrative fairness and predictability, reinforcing the need for algorithmic selection before deploying systems for public-fund allocation.

Figure 3 presents the mean F_1-macro and their corresponding standard deviations for each algorithm, aggregated across all training data fractions and text representation strategies. The Random Forest model achieved the highest nominal score (0.755 ± 0.003), closely followed by the Decision Tree (0.742 ± 0.004). Beyond their high predictive accuracy, both models demonstrate remarkable stability, as evidenced by their exceptionally narrow standard deviations. In contrast, linear and centroid-based models, such as Linear SVC and Nearest Centroid, clustered in an intermediate performance tier, with F_1 ranging from 0.55 to 0.59. The SGD model not only yielded the lowest average performance (0.495) but also exhibited the highest volatility among all evaluated techniques (± 0.094), indicating an excessive sensitivity to variations in the training data composition.

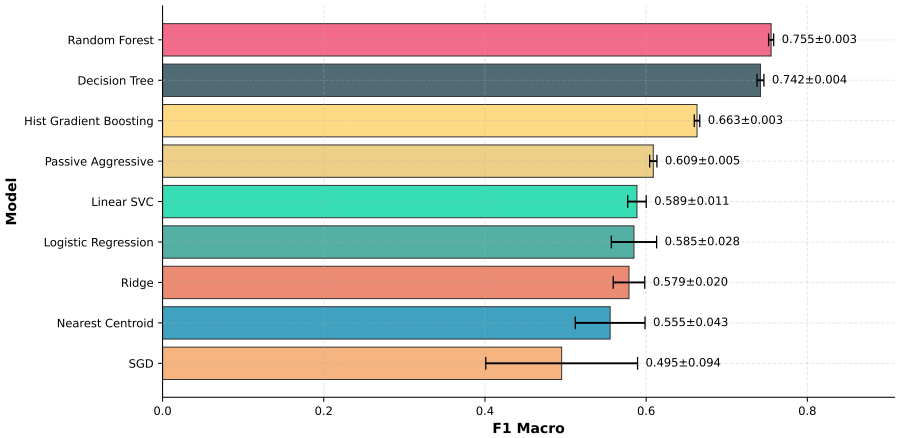


Figure 3. Macro F_1-macro and standard deviation across the nine classifiers on the hold-out test set.

Figure 4 shows that Random Forest, Decision Tree, and Histogram-based Gradient Boosting form the best-ranked group, with no statistically significant differences among them. The linear and centroid-based classifiers show substantial overlap, whereas SGD appears isolated at the lowest end of the ranking, reinforcing its inadequacy for this task. The integration of these predictive and statistical results provides fundamental insights

for the governance of artificial intelligence systems in the public sector. The statistical robustness and minimal variance demonstrated by the tree-based ensembles (Random Forest and Decision Tree) supply the necessary guarantees of predictability and administrative equity. For government resource allocation driven by AI, consistency is as vital as precision; a system that assigns drastically different adherence labels to the same academic production due to minor variations in the training set—as demonstrated by the SGD model—would be legally fragile and politically unsustainable. Therefore, the final model selection must prioritize the statistically stable top-tier cluster. This ensures that funding and strategic alignment decisions are based on robust, auditable technical merit rather than algorithmic instability, thereby shielding public administration from challenges regarding the impartiality of the automated decision-making process.

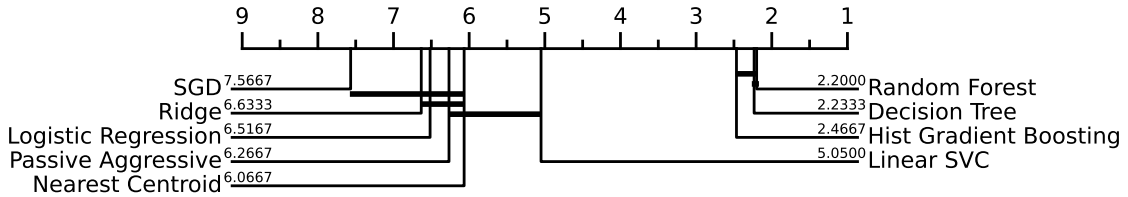
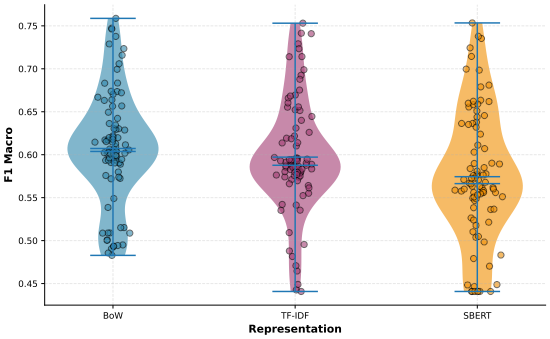


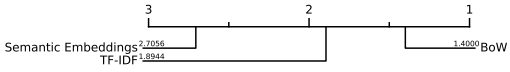
Figure 4. Critical Difference (CD) diagram based on the Friedman test followed by pairwise Wilcoxon signed-rank tests with Holm correction ($p < 0.05$). Connected classifiers are not statistically different; lower ranks indicate better performance.

Beyond algorithm selection, identifying the most effective feature-extraction strategy is crucial to the pipeline's overall efficiency. Figure 5(a) provides a distributional overview of the F_1-macro achieved across all evaluated models and data splits, categorized by their text representation: BoW, TF-IDF, and semantic embeddings. The violin plot reveals a counterintuitive phenomenon: the traditional sparse representations (BoW and TF-IDF) exhibit a higher concentration of performance mass in the upper quartiles than the dense semantic embeddings. While the maximum scores achieved are comparable across all three spaces—approaching 0.75 for the best-performing tree-based ensembles, the median performance for semantic embeddings is noticeably lower. Furthermore, semantic embeddings display an elongated lower tail, indicating that several classifiers struggled to optimize their decision boundaries within this dense, continuous space without specialized, algorithm-specific tuning.

Figure 5(b) shows that BoW achieves the best average rank (1.4000), closely followed by TF-IDF (1.8944). The horizontal line connecting them indicates no statistically significant difference between these sparse representations. In contrast, semantic embeddings obtain the worst average rank (2.7056) and appear isolated in the diagram, indicating inferior performance for this adherence classification task. Sparse representations such as BoW and TF-IDF are computationally inexpensive and require no specialized hardware, but they treat each token independently, fail to capture semantic relations between terms, and discard any token absent from the training vocabulary, making them more suitable for domains where lexical overlap between documents and categories is consistent. Semantic embeddings, in contrast, encode contextual meaning and cross-lingual relations, making them preferable when surface-level term matching is insufficient, but incur substantially higher computational costs.



(a) Violin plot



(b) CD diagram

Figure 5. Statistical analysis of the text representation strategies. (a) Performance distribution across all evaluated classifiers and data splits. (b) Critical Difference (CD) diagram based on the Friedman test followed by pairwise Wilcoxon signed-rank tests with Holm correction ($p < 0.05$).

Figure 6 presents the wall-clock training and inference times across all classifiers. Training times range from 0.33s for Nearest Centroid to 48.87s for Decision Tree, with Random Forest (8.73s) and Histogram-based Gradient Boosting (7.47s) occupying an intermediate tier. The inference time ordering, however, diverges considerably from training: Random Forest becomes the most expensive model at prediction time (0.1833s), followed by Histogram-based Gradient Boosting (0.0412s), while the remaining classifiers operate below 0.02s. Critically, these absolute values remain operationally negligible in the batch-processing scenario that characterizes government deployments — such as the periodic ingestion of new CAPES records — meaning that the final model selection can be driven entirely by predictive performance and stability criteria rather than computational constraints. This finding further consolidates the case for Random Forest as the recommended classifier, confirming that its superior F_1-macro comes at no meaningful infrastructure cost for public sector adoption.

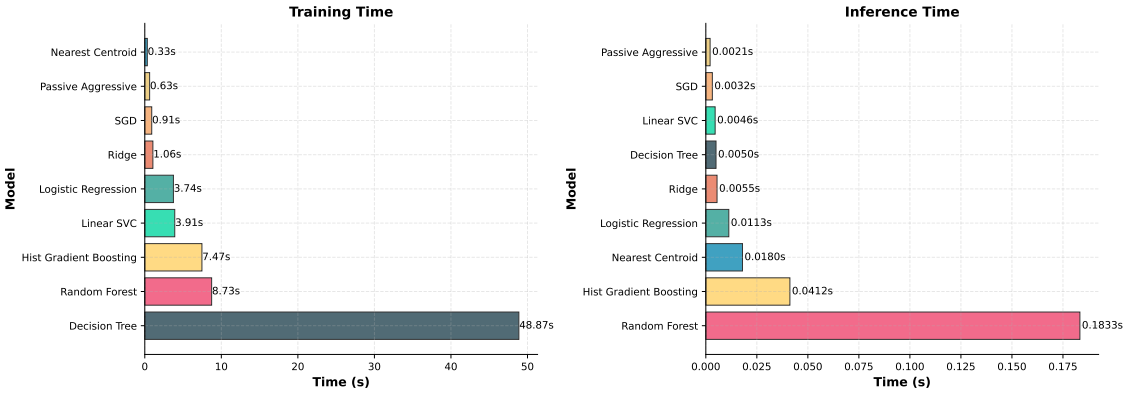


Figure 6. Training time (left) and inference time (right) for all evaluated classifiers, measured in seconds on the full training set and held-out test set, respectively.

These findings carry profound operational implications for public sector IT infrastructure. Advanced neural embedding architectures, such as semantic embeddings, generally require specialized hardware, such as GPUs, and significantly larger memory footprints for continuous, large-scale inference. In contrast, BoW and TF-IDF rely on

highly optimized sparse matrix operations that execute efficiently on standard, low-cost CPUs. Because empirical and statistical evidence categorically demonstrate that computationally expensive embeddings do not yield predictive gains and, in fact, degrade average algorithmic performance, the deployment of the BoW strategy is strongly recommended. This architectural choice ensures that government agencies can scale the automated classification of thousands of academic productions using existing computing infrastructure, maximizing the cost-effectiveness of the artificial intelligence system while preserving peak predictive precision.

Figure 7 illustrates the qualitative behavior of the generative module under correct and erroneous classification decisions. The evaluation of the fine-tuned PTT5-base model demonstrates a strong capability to synthesize contextually accurate explanations. Lexical overlap metrics show a robust ROUGE-1 of 0.6041, indicating that the generated texts successfully capture the primary vocabulary of the reference justifications. In open-ended natural language generation tasks, stricter exact-match metrics such as ROUGE-2 (0.3621), ROUGE-L (0.3888), and SacreBLEU (0.2442) exhibit lower absolute values, reflecting the model's capacity to paraphrase and generate alternative sentence structures.

The system's performance is particularly evident in a Semantic Similarity score of 0.9023, confirming that, despite variations in exact phrasing, the generated justifications preserve the core reasoning and semantics of the ground truth, thereby fulfilling the transparency requirements essential for evidence-based policymaking. In the success case, PTT5 accurately identifies the thematic alignment between educational robotics and the strategic theme of ICT for Education, reproducing key concepts such as teacher training and digital literacy in a coherent and contextually faithful justification. In the error case, the classifier assigns high adherence between a dissertation on sustainable urban drainage and the theme of Urban Mobility and Smart Cities, and the generative module faithfully follows this erroneous decision, producing a fluent justification grounded in superficial lexical overlap around terms such as "urban sustainability" and "urban planning".

6. Conclusion

We addressed a concrete governance challenge: enabling public managers to systematically assess the adherence of Brazilian postgraduate productions to state-level strategic themes defined under the CAPES National Agenda, at a scale that manual processes cannot sustain. Thus, we propose a pipeline that combines lightweight adherence classification with natural-language justification and evaluate it on 42k researcher-theme pairs.

The experimental results demonstrate that computationally lightweight models can serve as effective approximators of an LLM oracle, preserving satisfactory predictive fidelity without resorting to large-scale inference. Specifically, sparse representations paired with Random Forest achieved a F_1-macro of 0.755, statistically outperforming both denser embedding strategies and alternative classifiers while requiring negligible CPU inference time. This convergence of predictive quality and lightweight computation is particularly meaningful for the public sector, where infrastructure constraints are pervasive, and the total cost of deployment is a first-class concern. Beyond classification, a fine-tuned PTT5-Based model demonstrated that lightweight sequence-to-sequence architectures can generate explicit natural-language justifications conditioned on predicted labels, achieving a Semantic Similarity of 0.90 against human references and fulfilling

Correct Prediction (Prediction: High | Real label: High)

Human Reference: O título da dissertação 'ROBÓTICA EDUCACIONAL PARA FORMAÇÃO DE PROFESSORES DO CURSO TÉCNICO EM AGROPECUÁRIA' e as palavras-chave 'EDUCATIONAL ROBOTICS' e 'AUTOMATED IRRIGATION' indicam uma clara proximidade com o tema estratégico 'Tecnologias da Informação e Comunicação para Educação', especialmente com os descritores 'tecnologia na educação básica' e 'inovação educacional'. O resumo detalha o uso da robótica educacional como recurso pedagógico no ensino técnico em agropecuária, alinhando-se com a ideia de inovação educacional e aplicação de tecnologias na educação. A ênfase na formação de professores e no desenvolvimento do pensamento computacional também se conecta com os descritores 'formação de professores' e 'letramento digital'.
PTT5: Os dados do pesquisador mostram uma clara afinidade com o tema estratégico "Tecnologias da Informação e Comunicação para Educação". O título da dissertação, "ROBÓTICA EDUCACIONAL PARA FORMAÇÃO DE PROFESSORES DO CURSO TÉCNICO EM AGROPECUÁRIA", e o resumo destacam o uso de tecnologias educacionais, como robótica educacional, para a formação de professores, o que se alinha diretamente com a palavra-chave "tecnologia na educação básica" do tema estratégico. Além disso, o foco em "formação continuada" e "inovação educacional" está em consonância com a palavra-chave "formação continuada" do tema estratégico. A pesquisa também aborda a "inovação educacional" ao propor um "curso técnico integrado em agropecuária", o que se relaciona com "educação à distância" e "inclusão digital". Portanto, a afinidade é classificada como ALTA devido à correspondência explícita com elementos específicos do tema e dos descritores mais específicos.

Incorrect Prediction (Prediction: High | Real label: Low)

Human Reference: O trabalho do(a) pesquisador(a) aborda a drenagem urbana sustentável e a criação de espaços multifuncionais para mitigação de cheias, utilizando infraestrutura verde e azul. Embora o foco principal seja a gestão de águas pluviais e a resiliência urbana, há uma correlação temática com o tema estratégico de 'Recursos Hídricos', especialmente em relação à gestão de recursos hídricos e à conservação de recursos naturais. A menção à 'PIRAQUÊ-CABUÇU WATERSHED' e 'BLUE-GREEN INFRASTRUCTURE' indica uma conexão com bacias hidrográficas e soluções sustentáveis para a água, que são aspectos relacionados ao tema. No entanto, não há uma correspondência direta com descritores mais específicos do tema, como 'despoluição das águas' ou 'tecnologias de aproveitamento'. Portanto, a afinidade é classificada como BAIXA, devido à correlação temática genérica.
PTT5: O tema estratégico 'Mobilidade Urbana, Logística e Cidades Inteligentes' possui uma correlação temática com os dados do pesquisador, especialmente no que diz respeito ao planejamento urbano e à sustentabilidade ecológica. O título da dissertação 'MATRIZ DE PRIORIDADES PARA PROJETOS URBANOS MULTIFUNCIONAIS ESTRUTURADOS PELA DRENAGEM URBANA SUSTENTÁVEL' e o resumo destacam a importância do manejo de águas pluviais urbanas, que se alinha com a ideia de 'mobilidade sustentável' e 'planejamento urbano' presentes nas palavras-chave do tema estratégico. Ainda, a pesquisa aborda a sustentabilidade ambiental, que é uma das palavras-chave do tema estratégico. Porém, não há uma correspondência explícita com descritores mais específicos do tema, como 'cidades inteligentes' ou 'inteligência artificial'. Então, a afinidade é classificada como MÉDIA pela proximidade semântica geral com o tema, mas sem correspondência direta com descritores específicos.

Figure 7. Qualitative comparison between a correct prediction and an erroneous one. The first box shows the correct prediction. The second shows the incorrect prediction. All boxes show the human reference and PPT5 generation.

the transparency requirements mandated by public governance regulations.

Future research will focus on mitigating the cascading failure phenomenon identified in the qualitative analysis, where upstream classification errors compel the generative module to produce unfounded justifications. Implementing a confidence-based rejection mechanism could prevent such hallucinations, further strengthening the system's factual reliability. Additionally, extending this lightweight framework to broader scientific outputs, such as patents and peer-reviewed articles, would enable a more comprehensive mapping of the national innovation ecosystem to support strategic government planning.

Acknowledgments

This study was partially funded by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) – Finance Code 001. Additional support was provided by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), under grant number 130041/2025-4. Support was also provided by the *Fundação de Apoio à Pesquisa, ao Ensino e à Cultura* (FAPEC), under grant number 23104.003594/2025-12.

References

Annepaka, Y. and Pakray, P. (2025). Large language models: a survey of their development, capabilities, and applications. *Knowledge and Information Systems*, 67(3):2967–3022.

Awasthy, P., Trivedi, A., Li, Y., Bornea, M., Cox, D., Daniels, A., Franz, M., Goodhart, G., Iyer, B., Kumar, V., et al. (2025). Granite embedding models. *arXiv preprint arXiv:2502.20204*.

Barbier, V. and Jeangirard, E. (2025). Mapping scientific communities at scale. *arXiv preprint arXiv:2501.10035*.

Brasil (2021). Lei nº 14.129, de 29 de março de 2021: Princípios, regras e instrumentos para o governo digital e para o aumento da eficiência pública. Diário Oficial da União. Accessed: 2025.

CAPES (2021). Evolução do SNPG no decênio do PNPG 2011–2020. Technical report, Comissão Especial de Acompanhamento do PNPG 2011–2020 / CAPES, Brasília. Accessed: 30 mar. 2026.

CAPES (2023). Plano nacional de pós-graduação – PNPG 2024–2028: versão preliminar para consulta pública. Technical report, Ministério da Educação / CAPES, Brasília. Accessed: 30 mar. 2026.

CAPES (2026). Portal do observatório da pós-graduação. Accessed: March 30, 2026.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Gôlo, M. P. S. and Utino, M. Y. R. (2025). One-class lightweight interpretable filtering for academic profiles and strategic themes affinity. In *SBAC-PADW*, pages 147–154. IEEE.

ICMC (2025). Parceria entre ufms e usp desenvolve projeto inédito da capes que mapeia teses e dissertações em temas estratégicos da pós-graduação nos estados brasileiros. Accessed: 29 mar. 2026.

Jiang, Y., Pang, P. C.-I., Wong, D., and Kan, H. Y. (2023). Natural language processing adoption in governments and future research directions: A systematic review. *Applied Sciences*, 13(22):12346.

Jin, Z. and Mihalcea, R. (2022). Natural language processing for policymaking. In *Handbook of computational social science for policy*, pages 141–162. Springer.

Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, 535(7612):457–458.

Miranda, A. C. D., Utino, M. Y. R., Gôlo, M. P. S., Dos Santos, M. A. A., and de Souza, M. C. (2025). Reducing costs in large-scale classification: A hybrid bert–llm strategy. In *SBAC-PADW*, pages 131–138. IEEE.

Rigo, S. and Assunção, M., editors (2025). *International Symposium on Computer Architecture and High Performance Computing Workshops SBAC-PADW*.

Sakiyama, K. M., Fujimoto, M. L., Santin, R. V., and Rezende, S. O. (2025). Leandl hpc challenge 2025: Applying large-scale model adaptation techniques. In *SBAC-PADW*, pages 139–146. IEEE.

Utino, M. Y. R. and Gôlo, M. P. S. (2025). Efficient filtering with bert embeddings for researcher–topic affinity prediction in hpc pipelines. In *SBAC-PADW*, pages 155–162. IEEE.