

A Heterogeneous Ensemble Meta-Model for Large-Scale Categorization of Portuguese Scientific Records in OasisBR

Sergio Antonio Andrade de Freitas[1(✉)], André Corrêa da Silva[2], Milena de Faria Silva[4], Renan Carneiro Batista[3], and Washington Luís Ribeiro de Carvalho Segundo[3]

[1] CEDIS, University of Brasília, Brasília, Brazil
sergiofreitas@unb.br
[2] Banco do Brasil, Brasília, Brazil
[3] IBICT - Brazilian Institute of Information in Science and Technology, Brasília, Brazil
{renanbatista,washingtonsegundo}@ibict.br
[4] DAC – UNICAMP, Campinas, Brazil
m299743@dac.unicamp.br

Abstract. Large multidisciplinary repositories require reliable topical categorization to support retrieval, analytics, curation, and downstream research evaluation. This paper revisits and reconstructs an industrial research effort aimed at categorizing approximately 1.6 million Portuguese scientific records from OasisBR according to the nine CAPES large research areas. The operational snapshot preserved two critical assets: a large pre-labeled subset of roughly 800,000 records and aggregated benchmark results for several machine-learning algorithms. Building on these materials, this paper formalizes the task as a large-scale scientific-record classification problem and proposes a heterogeneous ensemble meta-model that combines Random Forest, XGBoost, CatBoost, support vector classification, logistic regression, Gaussian naive Bayes, decision trees, and AdaBoost through performance-informed weighted voting. The reconstructed pipeline integrates metadata ingestion, lexical normalization, and feature consolidation, reducing the raw lexical space from approximately 200,000 surface forms to 3,140 predictive features. In the harmonized benchmark, Random Forest achieved the highest mean F1-score across the nine large areas (0.7073), followed closely by XGBoost (0.7042). At the area level, Random Forest was the best standalone learner in seven areas and XGBoost in two. More importantly, in the operational unseen-batch evaluation, the ensemble meta-model outperformed the best standalone baseline in all nine areas, increasing the average hit rate from 19.86% to 33.30%, an average gain of 13.43% points. The strongest gains were observed in Engineering (+18.08 points), Human Sciences (+17.04), and Biological Sciences (+15.91). The main contribution of this work is a reproducible computational fram-

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 228–242, 2027.
https://doi.org/10.1007/978-3-032-30494-0_15

ing of large-scale scientific-record categorization for Portuguese metadata repositories, together with empirical evidence that heterogeneous ensembles offer more robust performance than single-model deployment in a difficult real-world setting.

Keywords: Text classification · Ensemble learning · Scientific databases · Information retrieval · Portuguese NLP · OasisBR

1 Introduction

The growth of scientific repositories has intensified the need for reliable large-scale categorization pipelines. In digital-library environments, categorization is not merely a convenience layer for browsing. It affects search quality, visibility of scientific production, statistical analyses, thematic monitoring, and the quality of downstream metadata services. Automated text categorization become a problem at the intersection of information retrieval, natural language processing, and scientific-database engineering [11,15,16].

This problem is pertinent in the Brazilian context. OasisBR, maintained by the Brazilian Institute of Information in Science and Technology, brings together open-access scientific production and research data from journals, institutional repositories, and digital libraries, thereby constituting an access point to articles, theses, dissertations, conference papers, datasets, preprints, and other scholarly objects [8]. Because the portal aggregates metadata from different institutions, record quality and topical consistency constitute operational requirements for the portal. Studies on metadata standardization in OasisBR indicate that interoperability and record normalization constitute conditions for information visibility and access in the portal [12].

A second domain requirement concerns the classification system adopted by CAPES. The CAPES table organizes knowledge into nine broad areas and subordinate hierarchical levels, providing a taxonomy used in research and graduate-education evaluation processes [4]. In repositories oriented toward scientific information retrieval and policy support, aligning metadata with this taxonomy favors record aggregation, thematic analysis, and the identification of related scholarly production.

This article addresses the problem of automatically categorizing OasisBR records into CAPES broad areas by means of machine learning. The project that originated this study considered a corpus of approximately 1.6 million records, of which about 800,000 already contained assigned categories and were used as training evidence. Although the initial version of the study presented empirical results, it did not formalize the computational workflow, the baseline definition, or the meta-model logic at the level required by contemporary LNCS papers. The present article reconstructs the study as a computational contribution grounded in reproducible procedures.

The central contribution consists in reformulating the generic notion of a "meta-model" as a heterogeneous ensemble meta-model. Instead of relying on

a single classifier, the proposed system combines distinct base models whose performance varies according to the lexical profiles of the records and the research areas considered. This design is motivated by the well-known observation that scientific texts exhibit heterogeneous distributions, domain-specific jargon, and area-dependent class boundaries, which makes single-model deployment brittle in practice [1,9,10,15]. Ensemble learning is a natural candidate for this setting because it can integrate complementary decision patterns while reducing the risk of over-specialization [2,6,14].

The contributions of this paper are fourfold:

1. it reframes the OasisBR categorization task as a large-scale scientific-record classification problem grounded in information-retrieval and scientific-database requirements;
2. it formalizes a heterogeneous ensemble meta-model based on performance-informed weighted voting over eight machine-learning learners;
3. it consolidates the preserved benchmark evidence into a coherent two-stage evaluation composed of a harmonized model-selection benchmark and an operational unseen-batch comparison; and
4. it identifies practical lessons for large multilingual or Portuguese-language repositories, especially regarding lexical consolidation, model robustness, and deployment-oriented evaluation.

The remainder of the paper is structured as follows. Section 2 discusses related work. Section 3 presents the problem setting and dataset context. Section 4 describes the proposed methodology. Section 5 reports the results. Section 6 discusses implications, limitations, and future work. Threats to validity are summarized in Sect. 7, and Sect. 8 concludes the paper.

2 Related Work

2.1 Scientific-Document and Research-Article Classification

Automated text categorization has a long research tradition, with classical surveys already noting the importance of representation choices, feature selection, and algorithmic bias toward particular data distributions [15,16]. More recent work has specialized this agenda toward scientific and scholarly corpora. Lu et al. proposed an ensemble approach for identifying artificial-intelligence articles in the Web of Science, combining rule-based filtering, SciBERT, decision trees, and support vector machines to improve recall and F1-score over keyword-only retrieval [10]. Lee et al. used pre-trained BERT to classify scientific articles about electromagnetic-field health risks into research categories, showing that title-and-abstract classification can support the curation of specialized scientific corpora [9]. These studies confirm two points that are central to the present work: first, scientific texts can be classified effectively from metadata fields such as title and abstract; second, ensemble or hybrid approaches are especially attractive in research-domain settings with fuzzy boundaries.

Related evidence also appears outside scholarly metadata repositories. Aubaid et al. compared machine-learning and rule-based embedding techniques for document classification and highlighted the role of representation and preprocessing choices in final performance [1]. Rahman et al. evaluated several ensemble configurations for comment classification and showed that heterogeneous combinations of learners can outperform homogeneous or single-model baselines [14].

2.2 Ensemble Learning for Text Classification

In machine learning, ensemble methods combine predictors with the objective of increasing generalization capacity and reducing sensitivity to variation in the data. Bagging and boosting constitute reference strategies in this literature [2,6]. Random Forest combines decision trees estimated from data resampling and from attribute subsets selected through randomization, a procedure suited to representations with many dimensions and sparsity [2]. Families of gradient boosting, such as XGBoost and CatBoost, extend the logic of boosting through training with regularization and support for scalability, with applications in tabular and textual representations [3,13]. Support vector classifiers retain relevance in high-dimensional feature spaces because of their margin-based formulation [5]. Logistic regression, in turn, provides linear decision boundaries and probability estimates in text-classification tasks [7].

The literature on text classification indicates that no learner achieves superior performance across all corpora and all label structures [15]. This limitation is relevant in tasks involving records of scholarly production, intersections among fields of knowledge, distinct metadata sources, and terminological variation. In such settings, ensemble learning can mitigate the fragility of individual models through the aggregation of decision patterns. The methodological problem consists of selecting base learners and defining a combination rule capable of preserving stability during system use.

2.3 Repository Engineering and Metadata Context

OasisBR is not a document collection, but a federated service for scientific information that aggregates objects from journals, repositories, and digital libraries [8]. In this type of infrastructure, classification quality depends on the quality of the available metadata. Marques et al. argue that metadata standardization constitutes a requirement for dissemination, interoperability, and access in the OasisBR portal [12]. This finding has consequences for machine learning, since category assignment depends on the consistency of fields such as title, abstract, keywords, and language tagging.

The CAPES taxonomy introduces an additional constraint to the problem. Because it is used to organize the evaluation of research and graduate education in Brazil, associating repository records with CAPES areas can support reporting, indexing, thematic analysis, and retrieval of related scholarly production [4]. The present paper therefore sits at the interface of machine learning and scientific-database engineering rather than in pure algorithm design.

3 Problem Formulation and Dataset Context

3.1 Task Definition

The task addressed in this paper is the automatic categorization of OasisBR scientific records into the nine CAPES large areas: Agricultural Sciences, Biological Sciences, Health Sciences, Exact and Earth Sciences, Human Sciences, Applied Social Sciences, Engineering, Linguistics, Literature and Arts, and Multidisciplinary. The article is based on a project snapshot in which the working repository contained approximately 1.6 million records, with roughly half of them already categorized and available as labeled evidence.

At a practical level, the input for each record is composed of metadata fields that are common in scientific repositories, especially title, abstract, and keywords. The output is a CAPES large-area label. From an information-systems perspective, the objective is to support retrieval, repository curation, and large-scale metadata enrichment. From a machine-learning perspective, the challenge lies in dealing with Portuguese scientific language, heterogeneous terminology, and variable metadata quality across areas.

3.2 Preserved Project Evidence and Reconstruction Strategy

The project materials preserved three classes of evidence: (i) a narrative description of the pipeline, (ii) aggregated benchmark indicators for multiple algorithms across the nine large areas, and (iii) an operational comparison between a single-model baseline and an ensemble meta-model on unseen document batches. However, the archived materials did not preserve all experiment scripts, fold-level traces, or a fully specified reproducibility package. To maximize scientific clarity while remaining faithful to the preserved evidence, this paper reconstructs the study around a conservative two-stage evaluation:

1. a harmonized benchmark used for model comparison and baseline selection across the nine large areas; and
2. an operational unseen-batch evaluation used to compare the selected standalone baseline against the proposed ensemble meta-model.

This reconstruction strategy avoids overclaiming unavailable details while still providing a coherent and reproducible computational framing.

3.3 The Feature Space

In revisiting the materials, we retain the observation that natural-language preprocessing and feature consolidation reduced the raw lexical space from approximately 200,000 surface forms to 3,140 predictive features. Although the historical version of the study does not specify all preprocessing procedures in detail, we interpret this reduction as consistent with a computational pipeline comprising metadata normalization, token filtering, stopword removal, and vocabulary consolidation. For Portuguese scientific corpora, such consolidation is methodologically important, since inflectional variation, multiword terminology, and metadata heterogeneity tend to increase representational sparsity.

4 Methodology

4.1 Overview

Figure 1 summarizes the proposed pipeline. The process begins with metadata ingestion from OasisBR records. The title, abstract, and keywords are normalized and consolidated into a predictive representation. Multiple base learners are then trained and evaluated. Finally, their outputs are integrated by a performance-informed ensemble meta-model.

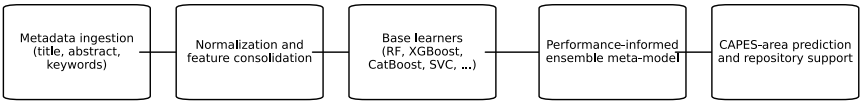

Fig. 1. Overview of the reconstructed categorization pipeline.

4.2 Metadata Preprocessing and Feature Engineering

The preprocessing stage is designed to stabilize the lexical representation of Portuguese scientific metadata. The preserved project description indicates that the final predictive space was obtained after aggressive lexical consolidation, reducing the raw space to 3,140 features. In the reconstructed formulation, the preprocessing stage comprises the following operations:

1. lowercasing and punctuation normalization for title, abstract, and keyword fields;
2. removal of obviously non-informative symbols and duplicated lexical fragments;
3. Portuguese stopword filtering and suppression of low-information tokens;
4. consolidation of morphological and near-duplicate variants into a reduced predictive vocabulary; and
5. document-vector construction over the consolidated lexical space.

The purpose of this stage is not only dimensionality reduction. It also reduces noise caused by repository heterogeneity. In scientific databases, records often originate from different institutional practices and export conventions. Vocabulary consolidation therefore acts as both a machine-learning step and a metadata-engineering step.

4.3 Base Learners

Eight base learners are considered in the harmonized benchmark: AdaBoost, CatBoost, Decision Tree, Gaussian Naive Bayes, Logistic Regression, Random Forest, Support Vector Classification (SVC), and XGBoost. These learners were selected because they cover distinct inductive biases and model families:

- tree ensembles and boosting methods are effective when complex non-linear decision patterns exist [2,3,6,13];
- margin-based classifiers are traditionally competitive in sparse text spaces [5]; and
- linear probabilistic models provide a stable reference point for high-dimensional text categorization [7,15].

The intent is not to claim that these eight learners exhaust the search space. Rather, they represent a sufficiently heterogeneous committee for robust ensemble construction.

4.4 Standalone Baseline Selection

The preserved benchmark reports per-area accuracy, precision, recall, and F1-score for several algorithms. From these results, the best standalone learner is selected for each large area according to F1-score. Table 1 shows the selected area-wise baselines.

Table 1. Best standalone learner by CAPES large area in the harmonized benchmark

Large area	Best model	Acc.	Prec.	Rec.	F1
Agricultural Sciences	Random Forest	0.679	0.821	0.679	0.706
Biological Sciences	Random Forest	0.654	0.852	0.654	0.709
Health Sciences	Random Forest	0.652	0.840	0.652	0.697
Exact and Earth Sciences	XGBoost	0.654	0.826	0.654	0.694
Human Sciences	XGBoost	0.670	0.813	0.670	0.705
Applied Social Sciences	Random Forest	0.696	0.897	0.696	0.753
Engineering	Random Forest	0.623	0.832	0.623	0.678
Linguistics, Literature and Arts	Random Forest	0.712	0.772	0.712	0.716
Multidisciplinary	Random Forest	0.700	0.838	0.700	0.721
Best-model frequency		Random Forest in 7 areas; XGBoost in 2 areas			

This area-wise best learner forms the simple-model baseline later used in the operational comparison.

4.5 Performance-Informed Ensemble Meta-Model

The reconstructed meta-model aggregates the class-probability outputs of all base learners through weighted voting. Let $p_m(c \mid d)$ denote the probability assigned by learner m to class c for document d. We define a global performance weight w_m from the harmonized benchmark as

$$w_m = \frac{\overline{F1}_m}{\sum_{j=1}^{M} \overline{F1}_j}, \tag{1}$$

where $\overline{F1}_m$ is the mean F1-score of learner m across the nine large areas. The meta-model score is then

$$S(c \mid d) = \sum_{m=1}^{M} w_m \, p_m(c \mid d), \tag{2}$$

and the final prediction is

$$\hat{y}(d) = \arg\max_{c} S(c \mid d). \tag{3}$$

This design is intentionally conservative. It does not require training an additional black-box stacker, yet it still exploits the relative strengths of stronger learners. Table 2 reports the mean benchmark performance used to derive the weights.

Table 2. Mean benchmark performance by algorithm across the nine large areas

Algorithm	Acc.	Prec.	Rec.	F1
Random Forest	0.670	0.833	0.670	0.707
XGBoost	0.667	0.826	0.667	0.704
CatBoost	0.660	0.816	0.660	0.696
SVC	0.659	0.822	0.659	0.696
Logistic Regression	0.638	0.790	0.638	0.670
GaussianNB	0.503	0.647	0.503	0.499
Decision Tree	0.485	0.581	0.485	0.499
AdaBoost	0.469	0.596	0.469	0.485

The weight distribution is therefore naturally dominated by Random Forest and XGBoost, but the ensemble still benefits from the complementary behavior of CatBoost, SVC, and Logistic Regression.

4.6 Evaluation Protocol

We organized the evaluation in two settings.

Harmonized benchmark. The first setting uses the preserved labeled evidence to compare base learners and select the area-wise best standalone baseline. The reported metrics are accuracy, precision, recall, and F1-score.

Operational unseen-batch evaluation. The second setting compares the simple-model baseline with the proposed meta-model on batches of previously unseen documents sampled from each large area. We treat this setting as a stricter and more deployment-oriented evaluation scenario than the harmonized benchmark. It therefore approximates the conditions faced by a repository service

when assigning categories to newly ingested records. At this stage, we retain two indicators: the number of exact hits and the hit rate for each broad area.

We adopt this two-stage protocol to support a more precise interpretation of the results. Benchmarks based on labeled and curated subsets allow us to estimate the relative quality of the models under controlled conditions. Operational batches, however, tend to exhibit greater noise, distribution shift, metadata sparsity, and topical ambiguity. Thus, a model that achieves competitive F1 performance in the benchmark may not preserve the same behavior under deployment conditions. We motivate the proposed meta-model from this gap between experimental performance and operational performance.

5 Results

5.1 Benchmark for Standalone Learners

In the harmonized benchmark, we observe a clear performance hierarchy among the base learners. Random Forest achieved the highest mean F1-score across the nine large areas (0.707), followed by XGBoost (0.704), CatBoost (0.696), and SVC (0.696). Logistic Regression remained competitive, although below this leading group. Gaussian Naive Bayes, Decision Tree, and AdaBoost formed a lower-performing cluster, with mean F1-scores close to or below 0.50. Figure 2 summarizes this ranking.

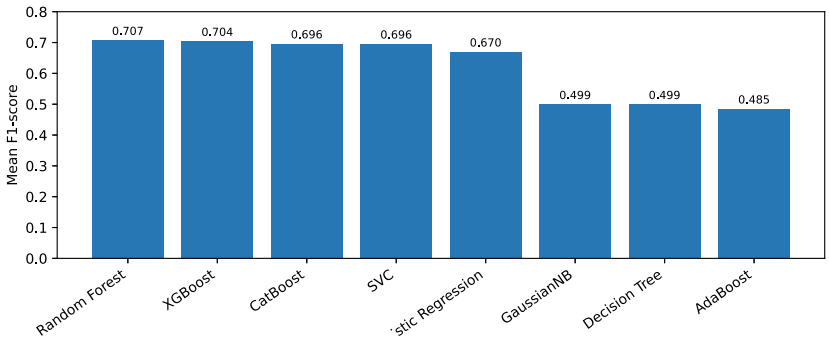

Fig. 2. Mean F1-score by algorithm in the harmonized benchmark.

We noted that two points deserve emphasis. First, tree-based ensembles dominate the ranking, which is consistent with their capacity to model non-linear feature interactions and to handle heterogeneous document distributions. Second, the small performance gap between Random Forest and XGBoost suggests that both models capture relevant, but not identical, patterns in the data. This result motivates their combination in an ensemble, rather than the selection of a single classifier.

At the area level, Random Forest was the best standalone learner in seven of the nine large areas, while XGBoost led in Exact and Earth Sciences and Human Sciences. The highest area-specific F1-score was observed in Applied Social Sciences (0.753 with Random Forest), followed by Multidisciplinary (0.721) and Linguistics, Literature and Arts (0.716). The lowest value among the selected best models was found in Engineering (0.678), suggesting that this area may be more lexically heterogeneous or more difficult to separate using metadata-only inputs.

5.2 Operational Comparison Between the Simple Model and the Meta-Model

Table 3 reports the unseen-batch comparison between the selected standalone baseline and the ensemble meta-model. The meta-model improved the hit rate in all nine large areas.

Table 3. Operational unseen-batch comparison between the simple-model baseline and the proposed meta-model

Large area	N	Simple hits (%)	Meta hits (%)	Gain (pts)
Agricultural Sciences	1497	307 (20.51)	432 (28.86)	+8.35
Biological Sciences	2495	235 (9.42)	632 (25.33)	+15.91
Health Sciences	1497	313 (20.91)	528 (35.27)	+14.36
Exact and Earth Sciences	1996	515 (25.80)	796 (39.88)	+14.08
Human Sciences	1497	315 (21.04)	570 (38.08)	+17.04
Applied Social Sciences	1497	314 (20.98)	488 (32.60)	+11.62
Engineering	1996	362 (18.14)	723 (36.22)	+18.08
Linguistics, Literature and Arts	1996	595 (29.81)	738 (36.97)	+7.16
Multidisciplinary	1497	182 (12.16)	396 (26.45)	+14.29
Average	–	**19.86**	**33.30**	**+13.43**

Figure 3 shows the same pattern graphically. The gains are consistent and substantial.

We observed the largest gains in Engineering (+18.08 %), Human Sciences (+17.04 %), and Biological Sciences (+15.91%). These areas appear to benefit more from ensemble aggregation, as they encompass heterogeneous terminologies and partially overlapping vocabularies. By contrast, Linguistics, Literature, and Arts showed the highest baseline hit rate and the smallest gain, suggesting that its lexical signal may be more homogeneous or more stable.

5.3 Benchmark Performance and Operational Accuracy

Area-level F1-scores in the range of 0.68–0.75 may seem difficult to reconcile with operational hit rates between 25% and 40%. This difference reflects the

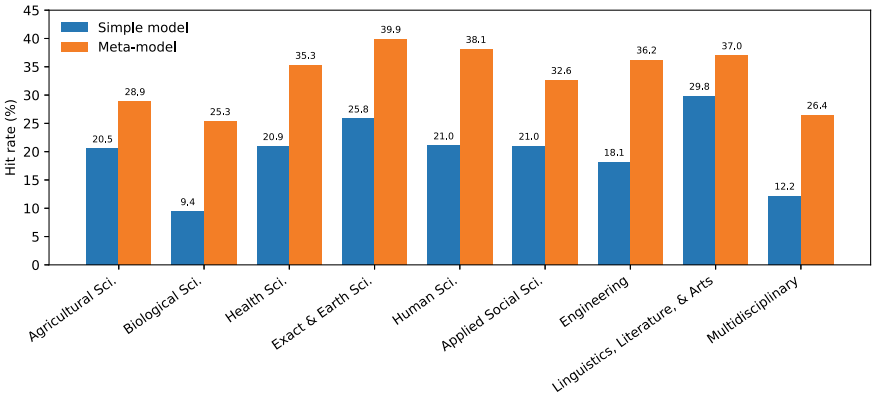

Fig. 3. Operational hit-rate comparison on unseen batches.

evaluation setting rather than an inconsistency in the results. The harmonized benchmark captures model behavior on curated labeled data, whereas the operational batches approximate deployment conditions. Records ingested into the repository may contain ambiguous metadata, limited abstracts, noisy keywords, or terminology associated with emerging research topics. Lower exact-hit rates are therefore expected in this setting.

This distinction is methodologically relevant for repository engineering. Evaluation in this domain should not be restricted to curated benchmark scores. Benchmarking supports the identification of candidate learners, but operational evaluation is required to assess whether the classification logic remains stable under input variability. The proposed meta-model addresses this requirement by combining models with complementary decision patterns and by reducing sensitivity to instability in the operational data.

6 Discussion

6.1 The Predominance of Tree-Based Ensembles

We interpret the predominance of Random Forest and XGBoost as consistent with the representational structure of scientific metadata. Titles, abstracts, and keywords encode lexical evidence together with interactions that vary across research areas and are not adequately captured by a strictly linear decision function. Tree-based ensembles are therefore appropriate for this setting, since they can model feature interactions without imposing an assumption of linear separability [2,3]. They also provide a suitable inductive bias for heterogeneous feature spaces, in which different subsets of terms may be informative for different areas of knowledge.

We relate this result to the heterogeneity of the predictive signal. Some records contain terminology that is strongly associated with a specific CAPES area, which makes their classification less ambiguous. Other records, however,

are shorter, interdisciplinary, weakly normalized, or semantically ambiguous. In this context, tree-based ensembles can reduce dependence on a single partition of the feature space. Random Forest does so by aggregating decisions from multiple trees estimated over resampled data and feature subsets, whereas XGBoost does so by sequentially fitting additive trees that correct residual errors. This helps explain why both methods performed better than simpler baseline models in the harmonized benchmark.

6.2 Area-Specific Gains of the Meta-Model

6.3 Area-Specific Gains of the Meta-Model

We observed the largest gains in Engineering, Human Sciences, and Biological Sciences. Although these areas differ in substantive scope, they share a common source of classification difficulty: internal lexical heterogeneity. Engineering records may contain specialized terminology distributed across several subfields. Human Sciences records often involve conceptual vocabularies and interdisciplinary intersections. Biological Sciences combine domain-specific terminology with expressions associated with health, environmental, and life-science topics. Under these conditions, we do not expect a single learner to provide uniform reliability across all lexical configurations. We address this limitation through a meta-model that combines classifiers with complementary inductive biases.

We interpret these gains as evidence that deployment-oriented categorization in large scientific repositories should prioritize robustness rather than optimization restricted to curated validation settings. From the perspective of repository operation, marginal improvements on a controlled validation subset are less consequential than reducing classification failures when new records are ingested. For this reason, we consider the average improvement of 13.43 % in operational hit rate to be one of the most relevant empirical findings of this study.

6.4 Implications for Scientific Databases and Information Retrieval

From the perspective of scientific-database engineering, our results indicate that topical classification should be understood as a repository-service layer rather than as an isolated machine-learning task. In large-scale repository infrastructures, improved categorization can enhance search filters, support thematic browsing, and provide more consistent analytical indicators for repository management.

The results also reinforce the methodological role of preprocessing. In federated repositories such as OasisBR, lexical normalization and metadata consolidation are not secondary implementation steps. They condition the quality of the feature space and directly affect the robustness of the classification models. Consequently, the performance of the proposed meta-model depends not only on the learning algorithms, but also on the capacity of the pipeline to reduce noise, sparsity, and inconsistency in the metadata.

A further implication concerns evaluation design. Scientific repositories continuously ingest records from multiple institutions and disciplinary communities,

which exposes classification systems to metadata drift and input heterogeneity. Models that perform well under static benchmark conditions may lose reliability in operational use. For this reason, we argue that repository-oriented evaluation should incorporate deployment-oriented scenarios. The two-stage protocol adopted in this study provides a pragmatic mechanism for assessing both controlled benchmark performance and operational behavior under more variable input conditions.

6.5 Future Methodological Extensions

The preserved project notes indicate a promising future direction based on synonym-aware feature engineering for Portuguese scientific text. Such an extension could reduce sparsity further by clustering semantically related surface forms, especially across institution-specific wording variants. Because this idea was not preserved under the harmonized benchmark, it is treated here as future work rather than as part of the evaluated contribution.

A second extension concerns modern transformer-based architectures. The related literature shows that BERT-family models can perform strongly on scientific article classification [9,10]. A natural continuation of this work is therefore a hybrid architecture in which transformer encoders provide semantic embeddings and the current ensemble acts either as a calibration or ranking layer.

7 Threats to Validity

Internal validity. The main threat concerns the archival nature of the preserved materials. Not all fold-level traces, hyperparameter logs, or operational scripts were retained. To mitigate this threat, the paper explicitly distinguishes between preserved evidence and reconstructed formalization, and it avoids unsupported causal claims.

Construct validity. The harmonized benchmark uses accuracy, precision, recall, and F1-score, while the operational stage reports exact-hit rates on unseen batches. These metrics capture related but not identical aspects of performance. We therefore interpret them as complementary rather than directly interchangeable.

External validity. The study focuses on Portuguese scientific metadata and on the CAPES taxonomy. Although the computational strategy is portable, the numerical results may not transfer directly to repositories with different metadata conventions, languages, or taxonomic systems.

Conclusion validity. The meta-model improved the operational hit rate in all nine areas, which is a strong empirical pattern. However, stronger statistical analysis would require fold-level or document-level prediction traces. The present paper therefore reports a robust comparative tendency rather than inferential significance testing.

8 Conclusion and Future Work

This paper reconstructed and strengthened a large-scale scientific-record categorization study for OasisBR and positioned it as a contribution to scientific databases, information retrieval, and applied machine learning. Using approximately 800,000 pre-labeled records from a working repository snapshot of about 1.6 million items, we formalized a heterogeneous ensemble meta-model for CAPES large-area categorization of Portuguese scientific metadata.

The empirical evidence supports three main conclusions. First, among the standalone learners, Random Forest and XGBoost are the most reliable choices, with Random Forest achieving the highest mean F1-score and dominating seven of the nine area-wise benchmarks. Second, a performance-informed ensemble meta-model improves robustness substantially in the harder operational setting, increasing the average hit rate from 19.86% to 33.30%. Third, metadata normalization and feature consolidation are decisive components of the pipeline, not marginal preprocessing details.

For repository operators, the results indicate that large-scale categorization of scientific records should combine heterogeneous ensemble methods, deployment-oriented evaluation protocols, and systematic metadata engineering. These components are not ancillary to the classification task; rather, they constitute conditions for the reliability of repository-scale categorization.

The study suggests some directions for further work. These include synonym-aware feature engineering, transformer-based embeddings for Portuguese scientific text, hierarchical classification beyond the CAPES large-area level, and continuous-learning pipelines capable of accommodating changes in repository content over time.

Acknowledgments. The authors acknowledge the project team involved in the original OasisBR categorization effort and the institutional collaborators who contributed to preserving the benchmark indicators used in this reconstruction. The authors also acknowledge the financial support provided by the Fundação de Apoio à Pesquisa do Distrito Federal (FAP-DF) through a research project.

References

1. Aubaid, A.M., Mishra, A., Mishra, A.: Machine learning and rule-based embedding techniques for classifying text documents. Int. J. Syst. Assur. Eng. Manag. **15**, 5637–5652 (2024). https://doi.org/10.1007/s13198-024-02555-w
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
3. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

4. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior: Tabela de Áreas de conhecimento/avaliação. CAPES official website (2024). Updated 2024-02-01. https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/instrumentos/documentos-de-apoio/tabela-de-areas-de-conhecimento-avaliacao. Accessed 07 Mar 2026
5. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). https://doi.org/10.1007/BF00994018
6. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997). https://doi.org/10.1006/jcss.1997.1504
7. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. Wiley, 3 edn. (2013). https://doi.org/10.1002/9781118548387
8. Instituto Brasileiro de Informação em Ciência e Tecnologia: O oasisbr. Oasisbr official website (2025). https://www.oasisbr.ibict.br/vufind/about/home. Accessed 07 Mar 2026
9. Lee, S.W., Kwon, J.H., Kim, D., Kim, E.J.: Research category classification of scientific articles on human health risks of electromagnetic fields using pre-trained BERT. ICT Express **10**(2), 336–341 (2024). https://doi.org/10.1016/j.icte.2023.08.006
10. Lu, M., Tang, L., Zhou, X.: An ensemble approach for research article classification: a case study in artificial intelligence. PeerJ Comput. Sci. **10**, e2521 (2024). https://doi.org/10.7717/peerj-cs.2521
11. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
12. Marques, G.S., Santos, B.E.D., Lima, M.C.M.D., Morais, C.T.D., Souza, M.G.D.: Padronização de metadados adotados no portal agregador de conteúdo científico oasisbr. Ciência da Informação 54(2) (2025). https://doi.org/10.18225/ci.inf.v54i2.7223
13. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: Unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. **31** (2018)
14. Rahman, M.M., Shiplu, A.I., Watanobe, Y.: Commentclass: a robust ensemble machine learning model for comment classification. Int. J. Comput. Intell. Syst. **17** (2024). https://doi.org/10.1007/s44196-024-00589-3
15. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002). https://doi.org/10.1145/505282.505283
16. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49 (1999). https://doi.org/10.1145/312624.312647