

Lexical profiles across institutional boundaries: a longitudinal analysis of Brazilian graduate research

Emerson Cordeiro Silva

State University of Santa Cruz (UESC)

Paulo Eduardo Ambrosio

State University of Santa Cruz (UESC)

Francisco Bruno Souza Oliveira

`fbsoliveira@uesc.br`

State University of Santa Cruz (UESC)

Research Article

Keywords: CAPES, disciplinary classification, lexical profiles, longitudinal validation, science mapping, theses and dissertations

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10906321/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Lexical profiles across institutional boundaries: a longitudinal analysis of Brazilian graduate research

Emerson Cordeiro Silva¹, Paulo Eduardo Ambrosio^{1,2},
Francisco Bruno Souza Oliveira^{1,2*}

¹Graduate Program in Computational Modeling in Science and
Technology (PPGMC), State University of Santa Cruz (UESC), Ilheus,
Bahia, Brazil.

²Department of Engineering and Computing, State University of Santa
Cruz (UESC), Ilheus, Bahia, Brazil.

*Corresponding author(s). E-mail(s): fbsoliveira@uesc.br;
Contributing authors: emersonsooo@yahoo.com.br; peambrosio@uesc.br;

Abstract

National science systems classify research into institutional fields, but the extent to which administrative boundaries coincide with textual structure remains an empirical question. We examine this relationship in Brazilian graduate research using the CAPES Catalogue of Theses and Dissertations and an interpretable lexical pertinence model reconstructed from Silva (2022). Two complementary term–area normalisations are combined to rank 49 harmonised Evaluation Areas. A strictly temporal design evaluates profiles built from 1987–2010 on 2011–2014 and profiles built from 1987–2014 on 2015–2018. Keywords provide the most temporally consistent representation, with Top-1 concordance of 50.95% and 49.24%, respectively. Frozen profiles built from 1987–2018 are then applied to 94,194 keyword-bearing documents from 2019: 49.69% coincide with the official Evaluation Area, 63.51% remain within the same CAPES Great Area or closer, and 79.18% remain within the same College or closer. Hierarchical analysis shows that divergence is heterogeneous: some areas are displaced mainly toward neighbouring fields, others remain within the same College, and others span multiple Colleges. Lexical dispersion measures further distinguish concentrated displacement from broad dispersion; for example, Teaching combines high hierarchical displacement with low entropy, whereas Interdisciplinary and Environmental Sciences show broad destination distributions. The results show that transparent lexical profiles can be used not only for longitudinal classification

but also to study the relationship between institutional taxonomies and the textual organisation of science without treating administrative labels as epistemic ground truth.

Keywords: CAPES, disciplinary classification, lexical profiles, longitudinal validation, science mapping, theses and dissertations

1 Introduction

Theses and dissertations are a distinctive source for studying the organisation of knowledge in a national graduate system. Each document records a research problem, textual descriptors, a graduate programme, an institution, a year of defence and an institutional disciplinary label. In Brazil, these labels are provided by the Coordination for the Improvement of Higher Education Personnel (CAPES), which organises graduate evaluation into Evaluation Areas, Great Areas and colleges. This makes the CAPES Catalogue of Theses and Dissertations a large longitudinal setting for examining how institutional classification relates to lexical evidence in scientific documents.

Disciplinary classification is a basic prerequisite for many bibliometric analyses, and the choice of classification scheme can itself affect evaluation results (Shu et al. 2020). At the document level, text-based methods provide one route for mapping research onto pre-existing disciplinary schemes (Goh et al. 2020). Importantly for the present study, organisational and content-based classifications need not coincide: substantial field-dependent differences in overlap have been documented when institutional affiliations are compared with cognitive classifications of publications (Guns et al. 2018). Systematic classification divergences can also reveal non-trivial relations among disciplines rather than functioning only as prediction errors (Lyutov et al. 2021). This perspective is particularly relevant when the objective is to study disciplinary boundaries rather than simply maximise predictive accuracy.

Within text classification, traditional term weighting schemes, especially TF-IDF, remain relevant because of their simplicity and interpretability (Salton and Buckley 1988). Supervised approaches can further incorporate class information directly into term weights (Debole and Sebastiani 2003; Lan et al. 2009; Chen et al. 2016; Dogan and Uysal 2020; Attieh and Tekli 2023). Titles, abstracts and keywords may also contain distinct disciplinary signals (Pech et al. 2022), and keyword occurrence patterns have been used to characterise disciplinary specialisation (Sampagnaro 2023). However, classification performance measured on contemporary random partitions may overstate generalisation when the corpus, vocabulary and institutional taxonomy change over time, a problem related to concept drift (Gama et al. 2014).

In the CAPES context, Silva (2022) proposed lexical pertinence profiles based on two term-area normalisations. One normalisation relates a term to its distribution across areas; the other relates the term to the lexical volume of each area. Their sum was used to classify academic documents by Evaluation Area. The method is

transparent and auditable, but its longitudinal generalisation had not been evaluated at national scale with temporally separated training and application windows.

This article asks whether simple lexical pertinence profiles constructed from historical CAPES records can represent and discriminate Evaluation Areas in later periods. It also asks how to interpret divergences between the institutional label and the lexical ranking: are they mostly distant errors, or do they remain close in the CAPES hierarchy? The goal is not to infer the “true” area of a thesis or dissertation, nor to audit CAPES decisions. Rather, the article studies the relationship between institutional classification and lexical proximity induced by historical profiles.

The contributions are fourfold. First, the article reconstructs the M_1 , M_2 and $ME = M_1 + M_2$ formulation of Silva (2022) in a longitudinal setting. Second, it harmonises CAPES Evaluation Areas into an experimental universe of 49 classes. Third, it evaluates the profiles in two independent temporal windows. Fourth, it applies the frozen historical profiles to the 2019 production and analyses divergences at the levels of Evaluation Area, Great Area, college and graduate programme.

2 Background

2.1 CAPES Evaluation Areas and the thesis catalogue

CAPES coordinates the evaluation of Brazilian graduate programmes. Its current structure groups Evaluation Areas into Great Areas and three colleges (CAPES 2024a,b). This hierarchy organises evaluation committees and criteria and, in the present analysis, provides an institutional scale for distinguishing exact agreement from progressively more distant divergences. It is not treated as an epistemic ground truth. The contemporary structure has 50 Evaluation Areas. The present study works with 49 harmonised historical classes because the experimental period and the available records do not map one-to-one onto the current taxonomy.

The CAPES Catalogue of Theses and Dissertations is publicly available through open datasets covering successive historical periods (CAPES 2026a,c,b). The database includes titles, abstracts, keywords, programme identifiers, institutions, years and Evaluation Areas. Because the study spans more than three decades, the target classes must be treated historically: names may change, areas may be split or reorganised, and institutional categories may not have a stationary meaning over the whole period.

2.2 Supervised term weighting and lexical profiles

In document classification, supervised term weighting aims to assign higher weights to terms that help discriminate classes. The literature includes several class-aware alternatives to unsupervised frequency-based schemes (Debole and Sebastiani 2003; Lan et al. 2009; Chen et al. 2016; Dogan and Uysal 2020; Attieh and Tekli 2023). The formulation used here belongs to this interpretable family: instead of learning a black-box model, it estimates a term-area profile from observed frequencies. This deliberately transparent representation differs from contemporary semantic classifiers: its purpose here is not to establish a state-of-the-art predictive benchmark, but to

retain an auditable link between observed terms, institutional areas and subsequent divergence patterns.

This simplicity is important for the application. A lexical profile can be inspected term by term and area by area. The resulting classification is therefore not only a prediction but also a descriptive representation of how terms are distributed in the historical corpus. This article uses that property to interpret divergences between the official CAPES label and the lexical ranking as evidence of proximity, ambiguity or dispersion in the institutional taxonomy.

2.3 Evaluation measures

The primary metric is Top-1 concordance, equivalent to accuracy in a single-label classification problem: a document is counted as concordant when the highest-ranked lexical area equals the official CAPES Evaluation Area. This measure is intuitive but restrictive because there are 49 possible classes and many areas are institutionally close.

Top- k concordance evaluates whether the official area appears among the k highest-ranked areas. Top-3 and Top-5 are useful here because a thesis or dissertation may have a lexical profile close to multiple neighbouring areas. Precision, recall and F1 are class-sensitive metrics. For a given area, recall measures the share of documents officially assigned to that area that are recovered by the classifier, while precision measures the share of documents predicted as that area that are officially assigned to it. F1 is the harmonic mean of precision and recall (Sokolova and Lapalme 2009). Macro-F1 averages F1 over classes and gives the same weight to small and large areas; weighted F1 weights each class by its support. These metrics complement Top-1 because the class distribution is highly heterogeneous.

In addition, this article introduces a hierarchical reading of divergence. When the predicted area differs from the official area, the pair is classified into four mutually exclusive cases: same Evaluation Area (AA), different area within the same Great Area (GA), different Great Area within the same college (CO), and different college (XC). Cumulative quantities are also reported: within the same Great Area includes AA and GA, and within the same college includes AA, GA and CO. These cumulative labels do not imply value judgements; they only express distance in the CAPES hierarchy.

3 Data and methods

3.1 Corpus and temporal design

The corpus consists of CAPES thesis and dissertation records from 1987 to 2019. The records are predominantly in Portuguese, as expected for theses and dissertations submitted to Brazilian graduate programmes, although records in other languages may occur in the CAPES database. No language-based exclusion criterion was applied. Three textual fields were processed separately: title (T), abstract (A) and keywords (K). The target variable is the harmonised CAPES Evaluation Area. The article stops in 2019: no documents from 2020 onwards are used in the analyses reported here.

Profile construction	Evaluation/application	Role
1987–2010	2011–2014	temporal validation 1
1987–2014	2015–2018	temporal validation 2
1987–2018	2019	final application

Fig. 1 Longitudinal design of the study. Each profile is built only from documents earlier than the corresponding evaluation or application period.

OpenAI ChatGPT was used to assist with computational workflow development and manuscript preparation. All methodological decisions, analyses, interpretations and the final manuscript were reviewed and approved by the authors.

The experimental design is strictly temporal. First, profiles estimated from 1987–2010 are evaluated on 2011–2014. Second, profiles estimated from 1987–2014 are evaluated on 2015–2018. Finally, profiles estimated from 1987–2018 are applied to all eligible 2019 documents. The first two stages are validation windows; the last stage is a final out-of-time application (temporal holdout) used to characterise the 2019 production. Documents from a test or application year are never used to build the corresponding profiles. Figure 1 summarises this chronology and makes the no-leakage constraint explicit: every evaluation or application period occurs strictly after the records used to construct its lexical profiles.

3.2 Lexical preparation

Titles (T), abstracts (A) and keywords (K) were analysed as separate token lists. Text fields were converted to strings, Unicode-normalised to NFC, conservatively checked for common encoding artefacts, and stripped of carriage returns, line breaks and tabs before repeated whitespace was collapsed. Tokenisation was then performed after lowercasing by replacing non-word characters with whitespace using a Unicode-aware regular expression. Accented characters were preserved. Given the predominantly Portuguese-language corpus, a fixed Portuguese stopword list was applied, and one-character tokens were removed. No stemming, lemmatisation, morphological normalisation, n-gram construction or minimum-frequency threshold was used. Numeric and alphanumeric tokens were retained because they may carry domain-specific information. The same preprocessing was applied independently to titles, abstracts and keywords in every temporal experiment. Because the purpose is to reproduce and evaluate the lexical pertinence formulation, the analysis remains at the level of term frequencies rather than semantic embeddings.

Post-processing audits examined rare, numeric, alphanumeric, anomalously long and potentially malformed tokens. Items provisionally identified as requiring correction or removal represented approximately 0.0031% and 0.0004% of corpus occurrences, respectively. Given their negligible prevalence and to avoid post-hoc modification of the historical pipeline, no additional heuristic filtering was introduced.

The 2019 application had 94,497 theses and dissertations with official Evaluation Areas and no new class absent from the 1987–2018 reference universe. After field-specific lexical coverage checks, 94,481 title records, 94,260 abstract records and 94,194 keyword records had at least one term present in the training vocabulary.

The original CAPES records contained 59 distinct historical Evaluation Area labels between 1987 and 2018. A conservative longitudinal harmonisation was applied only when programme continuity supported a nominal or taxonomic succession. Ten historical labels were thereby mapped to successor labels, producing an analytical space of 49 Evaluation Areas. Potential splits, newly created areas, reallocations among coexisting areas and transversal categories were not treated as simple equivalences. The same 49-area space covered all classes observed in the 2019 holdout, with no previously unseen Evaluation Area. The harmonisation therefore provides a temporally consistent analytical space and should not be interpreted as implying that the CAPES institutional taxonomy remained substantively unchanged over time. The audited mapping is provided in the Supplementary Information.

3.3 Lexical pertinence metrics

Let t denote a term in the vocabulary and a a CAPES Evaluation Area. Let a' index all areas in the training universe, and let $f(t, a)$ be the frequency of term t in documents officially associated with area a during the profile construction period. The first component is

$$M_1(t, a) = \frac{f(t, a)}{\sum_{a'} f(t, a')}. \quad (1)$$

The denominator is the total frequency of term t across all areas. Therefore, $M_1(t, a)$ is the share of occurrences of term t attributed to area a . It answers the question: among all occurrences of this term, how concentrated is it in this area?

Let t' index all terms observed in area a . The second component is

$$M_2(t, a) = \frac{f(t, a)}{\sum_{t'} f(t', a)}. \quad (2)$$

The denominator is the total lexical volume of area a in the training period. Thus, $M_2(t, a)$ expresses the relative contribution of term t to the vocabulary of area a . It answers a different question: within this area, how prominent is this term?

The main metric reconstructed from Silva (2022) is the sum

$$ME(t, a) = M_1(t, a) + M_2(t, a). \quad (3)$$

Here $ME(t, a)$ combines cross-area concentration and within-area prominence. For a document d , let $f_d(t)$ be the frequency of term t in that document. The area score is

$$S(d, a) = \sum_t f_d(t) ME(t, a), \quad (4)$$

where the sum is over the terms of document d that are present in the training vocabulary. Each document is assigned a full ranking of the 49 areas by decreasing $S(d, a)$.

The equal contribution of M_1 and M_2 was not selected by optimisation on either evaluation period; it preserves the original formulation proposed by Silva (2022). To

assess the contribution of each component without changing that formulation, M_1 and M_2 were also evaluated separately as internal ablations in both temporal validation windows. Optimisation of their relative weights was outside the scope of the present study.

3.4 Quantifying lexical dispersion and institutional distance

Flat concordance does not distinguish two different phenomena: a field may systematically diverge toward one dominant neighbouring area, or it may distribute its documents across many distinct lexical destinations. We therefore characterise the destination distribution of each official Evaluation Area. Let p_{aj} be the proportion of documents officially assigned to area a whose highest-ranked lexical destination is area j , and let $K = 49$ be the number of harmonised areas. Following the information-theoretic definition of entropy introduced by Shannon (1948), and its later use in diversity and interdisciplinarity measurement (Shannon 1948; Stirling 2007; Mutz 2022), normalised Shannon entropy is

$$H_N(a) = \frac{-\sum_{j=1}^K p_{aj} \log p_{aj}}{\log K}. \quad (5)$$

Here $H_N(a) \in [0, 1]$: values near zero indicate concentration in few destinations, whereas larger values indicate a more even distribution over the classification space. In diversity analysis, entropy primarily captures the balance of a distribution together with the number of occupied categories; it does not by itself encode how dissimilar those categories are (Stirling 2007; Leydesdorff et al. 2019; Mutz 2022). For interpretability, we also report the effective number of destinations, $N_{\text{eff}}(a) = \exp(H(a))$, where $H(a)$ is the unnormalised Shannon entropy.

To incorporate the institutional topology, we define a CAPES hierarchical distance $d(i, j)$ between two Evaluation Areas: 0 for the same area, 1 for different areas in the same Great Area, 2 for different Great Areas in the same college, and 3 for different colleges. The mean displacement of official area a is

$$D(a) = \sum_j p_{aj} d(a, j). \quad (6)$$

Thus, entropy measures the diversity and balance of lexical destinations, whereas $D(a)$ measures how far those destinations lie from the institutional origin.

Finally, motivated by the diversity framework that distinguishes variety, balance and disparity (Stirling 2007) and by Rao–Stirling applications in scientometrics (Leydesdorff et al. 2019; Mutz 2022), we combine destination diversity with disparity through a hierarchy-based Rao–Stirling index,

$$R(a) = \sum_i \sum_j p_{ai} p_{aj} \frac{d(i, j)}{3}. \quad (7)$$

The division by 3 normalises the maximum pairwise CAPES distance to one. We use these measures descriptively as statistics of lexical dispersion; they are not treated as direct measures of “true” interdisciplinarity. This distinction is essential because institutional interdisciplinarity, lexical heterogeneity and structural bridging are related but non-equivalent constructs.

3.5 Sensitivity to Evaluation Area size

Because the number of theses and dissertations differs substantially among Evaluation Areas, we assessed whether area-level patterns were associated with class size. For the 49 areas in the 2019 keyword application, Spearman rank correlations were computed between the number of documents in each official area and exact area concordance, normalised Shannon entropy, mean hierarchical displacement and hierarchy-based Rao–Stirling diversity. Spearman correlation was used because it does not require a linear relation or normally distributed area-level indicators. These associations are interpreted descriptively and not causally.

3.6 Lexical map of the 2019 application

For the 2019 keyword application, each classifiable thesis or dissertation is represented by its complete 49-dimensional vector of *ME* scores across the Evaluation Areas. The score vector was L1-normalised to unit sum for each document before dimensionality reduction, retaining the relative lexical evidence for all Evaluation Areas rather than only the Top-1 destination. The resulting 94,194-by-49 matrix was projected directly to two dimensions with Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018), using cosine distance, 50 neighbours, a minimum distance of 0.10 and a fixed random seed of 42. UMAP is used only as a visualisation device: local neighbourhoods can reveal structure in the lexical score space, but global distances and apparent cluster separations in the two-dimensional projection are not treated as quantitative evidence. All quantitative comparisons are based on the original classifications, the CAPES hierarchy and the dispersion statistics defined above. For display, only the extreme 0.1% tails of each UMAP coordinate were clipped from the plotting window; no documents were removed from the analyses.

3.7 Computational environment

The experiments were run under Ubuntu 24.04 LTS on WSL2, with Python 3.12.3 and DuckDB 1.5.5. The machine used an Intel Core Ultra 7 155H processor with 22 logical CPUs available to the Linux environment, 15 GiB of RAM allocated to WSL, 4 GiB of swap and a local 1 TB filesystem. No GPU was used. Auxiliary Python libraries were used for data handling and figure generation; the main aggregations were executed in DuckDB over Parquet files. Building the 1987–2018 profiles and applying them to 2019 took 1 min 50 s of wall-clock time, with a maximum resident set size of approximately 11.3 GiB.

Table 1 Temporal validation of $ME = M_1 + M_2$ by textual field. Values are percentages.

Field	Window	Top-1	Top-3	Top-5	Macro-F1
T	2011–2014	49.46	73.42	82.61	37.72
A	2011–2014	42.43	65.79	76.51	27.30
K	2011–2014	50.95	74.79	83.57	40.78
T	2015–2018	45.87	70.12	79.84	36.18
A	2015–2018	38.30	62.36	73.23	25.77
K	2015–2018	49.24	73.22	82.55	40.43

4 Results

4.1 Temporal validation

Table 1 shows that the lexical profiles generalise to later periods. Keywords were the strongest field in both validation windows. Their Top-1 concordance decreased only from 50.95% in 2011–2014 to 49.24% in 2015–2018. Title was intermediate, and the abstract representation was the weakest field under the original ME formulation.

As an internal ablation, the two components of the pertinence metric were also evaluated separately. For titles and keywords, the combined metric $ME = M_1 + M_2$ consistently produced slightly higher Top-1 concordance than either component alone in both temporal windows. For keywords, Top-1 values for M_1 , M_2 and ME were 50.79%, 36.09% and 50.95% in 2011–2014 and 49.08%, 35.72% and 49.24% in 2015–2018. The small gain of ME over M_1 for keywords indicates that Top-1 discrimination in this field is dominated by the cross-area concentration captured by M_1 . This does not make M_2 redundant: it measures a distinct property—the relative prominence of a term within an area—and its contribution varies with textual field and rank depth. Abstracts illustrate this difference, as M_2 produced higher Top-3 and Top-5 concordance in both periods and slightly higher Top-1 concordance in 2015–2018 (39.11% versus 38.30% for ME). The two components were not tuned to maximise performance; ME is retained because it is the original formulation of Silva (2022) whose temporal generalisation is examined here. Full ablation results are reported in the Supplementary Information.

Figure 2 highlights the same pattern visually. The keyword representation changes comparatively little across windows, while title and abstract show larger declines. This supports the use of keywords as the main field for the subsequent 2019 structural analysis, with title and abstract retained as robustness comparisons.

4.2 Application to 2019

Profiles built from 1987–2018 were applied to the 2019 theses and dissertations. Table 2 reports the field-level results. Keywords again produced the strongest performance: 49.69% Top-1, 74.32% Top-3 and 84.12% Top-5. The median rank of the official area was 2 for all three fields, indicating that even many non-Top-1 cases remain near the top of the lexical ranking.

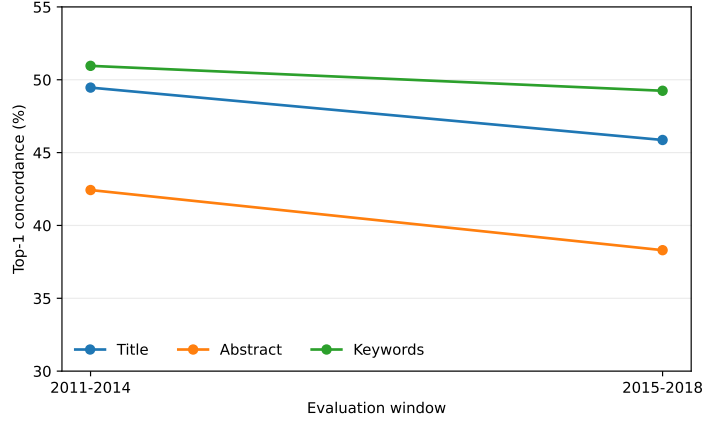


Fig. 2 Top-1 concordance in the two temporal validation windows. Profiles are always built from earlier years than the corresponding evaluation window.

Table 2 Final 2019 application using profiles built from 1987–2018.

Field	Documents	Top-1	Top-3	Top-5	Mean rank	Median rank
T	94,481	46.38	71.14	81.12	3.76	2
A	94,260	38.58	63.54	74.84	4.97	2
K	94,194	49.69	74.32	84.12	3.33	2

4.3 Lexical map and hierarchical structure in 2019

A flat Top-1 view hides part of the structure of the 2019 results. Figures 3 and 4 show that this structure is already visible at the aggregate level. The exclusive outcomes separate exact-area concordance from progressively more distant institutional divergences, whereas the cumulative view shows how much of the lexical classification remains within the same Great Area or College. Keywords combine the highest exact-area concordance (49.69%) with 63.51% of documents remaining within the same Great Area or closer and 79.18% within the same College or closer. Title and abstract show the same hierarchical ordering, although at lower exact concordance.

Figure 5 then shows the lexical score space itself. Each point represents one of the 94,194 keyword-bearing theses and dissertations; its position is determined by the complete 49-dimensional vector of L1-normalised *ME* scores, while colour in panel (a) represents the official CAPES College. The projection reveals substantial local organisation by institutional College, but also extensive overlap. Panel (b) isolates the 6,232 documents officially assigned to the Interdisciplinary Evaluation Area. Their lexical Top-1 destinations span all three Colleges: 53.87% are predicted in Humanities, 25.43% in Life Sciences and 20.70% in Exact, Technological and Multidisciplinary Sciences. More importantly, those documents reach all 49 Evaluation Areas as lexical Top-1 destinations. The pattern is therefore not merely a redistribution among the three Colleges but an exceptionally broad dispersion across the institutional classification space.

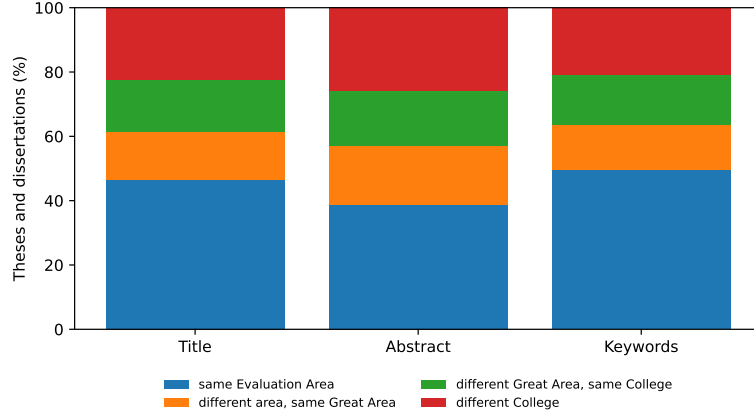


Fig. 3 Exclusive hierarchical outcomes in the 2019 application for title, abstract and keywords. The four categories are same Evaluation Area (AA), different area in the same Great Area (GA), different Great Area in the same College (CO), and different College (XC).

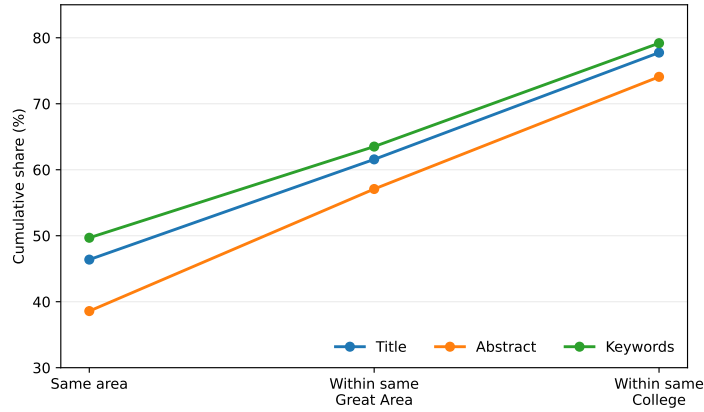


Fig. 4 Cumulative hierarchical concordance in the 2019 application. “Within same Great Area” includes AA and GA; “within same College” includes AA, GA and CO.

This interpretation is supported quantitatively below by the destination-diversity statistics rather than inferred from UMAP geometry alone; it should not be read as direct evidence of substantive interdisciplinarity.

Figure 6 disaggregates the same hierarchical structure by Evaluation Area and shows that the aggregate percentages conceal sharply different area-level patterns. Four mutually exclusive outcomes are reported: same Evaluation Area (AA), different area within the same Great Area (GA), different Great Area within the same college (CO), and different college (XC). Across all keyword-bearing documents, 49.69% are AA, 13.82% GA, 15.67% CO and 20.82% XC. Thus, 63.51% remain within the same Great Area or closer and 79.18% within the same college or closer.

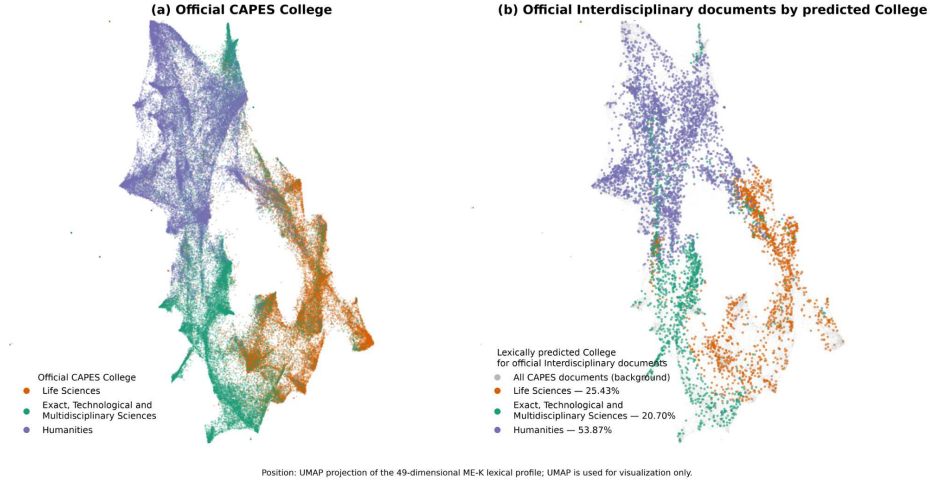


Fig. 5 Lexical map of the 2019 keyword application. Each point is a thesis or dissertation and its position is obtained by projecting the full 49-dimensional vector of L1-normalised *ME* scores directly with UMAP (cosine metric, $n_{\text{neighbors}} = 50$, $\text{min_dist} = 0.10$, random seed 42). (a) Points are coloured by the official CAPES college. (b) The grey background contains all 94,194 classifiable documents; coloured points are the 6,232 documents officially assigned to the Interdisciplinary Evaluation Area, coloured by the college of their lexical Top-1 destination. UMAP is used for visualisation only; quantitative interpretations are based on the original score space and CAPES hierarchy.

The area-level patterns show why XC should not be interpreted as a single type of classification error. Teaching has 83.82% XC, but its divergence is strongly concentrated: 78.90% of all Teaching documents are lexically assigned to the Humanities college and 71.33% specifically to Education. Biotechnology also has high XC (76.71%), yet 74.83% of its documents are assigned to Life Sciences, especially Agrarian Sciences I. Materials behaves differently: despite only 2.36% AA, 88.68% of its documents remain in the same college, mainly through Engineering II, Chemistry and Engineering III. Environmental Sciences is more transversal: its lexical destinations are distributed across Life Sciences (42.75%), Exact, Technological and Multidisciplinary Sciences (30.42%) and Humanities (26.83%). Interdisciplinary combines low self-retention (5.89%) with a broad cross-college distribution. These contrasts indicate that low exact concordance may reflect concentrated attraction to another disciplinary domain, redistribution among institutionally neighbouring fields, or genuinely broad lexical dispersion.

4.4 Lexical dispersion across Evaluation Areas

The dispersion statistics formalise this distinction. Among all 49 Evaluation Areas, Interdisciplinary has the highest normalised Shannon entropy ($H_N = 0.783$) and the highest Gini-Simpson diversity (0.923). Its 6,232 documents reach all 49 possible lexical destinations, corresponding to 21.07 effective destinations. Its mean hierarchical displacement is 2.670, second among the 49 areas, and its hierarchy-based Rao-Stirling diversity is 0.788, also second. Environmental Sciences has the highest Rao-Stirling value (0.796), while Biotechnology has the largest mean displacement (2.711).

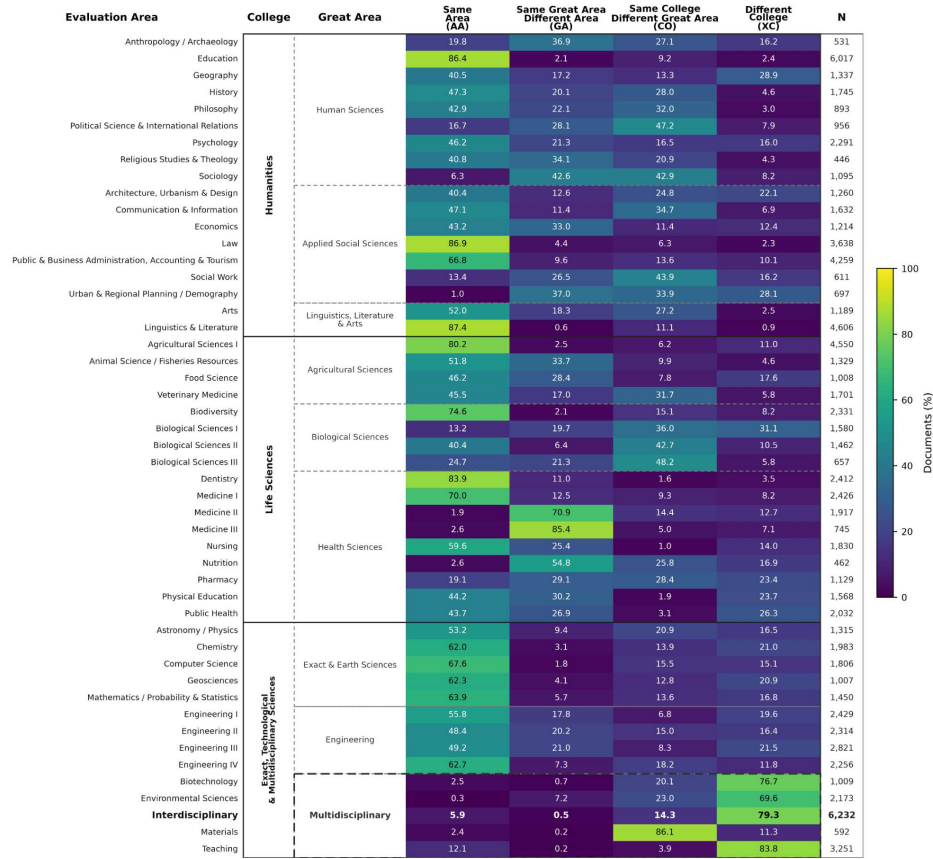


Fig. 6 Hierarchical concordance and divergence for the 49 CAPES Evaluation Areas in the 2019 keyword application ($N = 94,194$ classifiable theses and dissertations). Rows follow the institutional structure College $\rightarrow$ Great Area $\rightarrow$ Evaluation Area, with Evaluation Areas alphabetised within each Great Area. Percentages sum to 100% within each Evaluation Area. AA denotes the same Evaluation Area; GA, a different Evaluation Area in the same Great Area; CO, a different Great Area in the same College; and XC, a different College. The Interdisciplinary row is emphasised for discussion but is not treated as a separate hierarchical category. Multidisciplinary is a Great Area administratively located within the College of Exact, Technological and Multidisciplinary Sciences.

Teaching provides a useful counterexample: its mean hierarchical distance is 2.594 but normalised entropy is only 0.313 (rank 43), consistent with the strong concentration toward Education observed above. Materials has intermediate entropy (0.499; rank 22) despite high displacement (2.064). Lexical divergence and lexical dispersion are therefore empirically distinct dimensions. We use these measures descriptively and do not interpret lexical dispersion as a direct measure of substantive interdisciplinarity.

Area size was moderately associated with the observed lexical structure. Across the 49 areas, the number of 2019 keyword-bearing documents was positively associated with exact area concordance (Spearman $\rho = 0.563$, $p < 0.001$) and negatively associated with normalised entropy ($\rho = -0.482$, $p < 0.001$), mean hierarchical displacement

Table 3 Programme-level aggregation of the 2019 keyword application.

Group	Programmes	Docs.	Median n	Same area	Same Great Area	Same college
All programmes	4,253	94,194	18	47.06	69.23	90.00
$n \geq 10$	3,587	89,809	20	49.38	69.57	90.00

($\rho = -0.449$, $p = 0.001$) and hierarchy-based Rao–Stirling diversity ($\rho = -0.421$, $p = 0.003$). The direction of these associations is consistent with larger areas tending, on average, to exhibit more concentrated lexical signatures under this representation. This is a descriptive association rather than evidence that area size causes lexical consolidation. Size therefore contributes to area-level variation but does not determine it. Interdisciplinary and Education illustrate this clearly: despite similarly large supports ($N = 6,232$ and $N = 6,017$, respectively), their exact concordances are 5.89% and 86.36%, and their normalized entropies are 0.783 and 0.182.

4.5 Graduate programmes

Aggregating the keyword results by graduate programme provides a programme-level view without ranking individual programmes. In 2019, 4,253 programmes were identified. For the 3,587 programmes with at least ten documents, covering 89,809 theses and dissertations, the median exact concordance was 49.38%, the median share within the same Great Area was 69.57%, and the median share within the same college was 90.00% (Table 3). The interquartile range for exact concordance was wide, from 15.50% to 75.64%.

This aggregation is descriptive. Programme-level divergences may reflect networks, professional modalities, heterogeneous research lines, or later changes in the CAPES structure. They should not be read as normative evaluations of programme quality or adequacy.

5 Discussion

5.1 From document classification to a map of institutional–lexical relations

The longitudinal experiments provide evidence that the lexical model retains comparable performance across temporally separated windows before it is used descriptively. This order matters. A structural interpretation of 2019 divergences would be difficult to defend if the profiles collapsed under temporal separation. Instead, keyword Top-1 concordance remains close to 50% in both validation windows, while Top-5 remains above 82%. The 2019 application reproduces the same scale of performance. The resulting divergence matrix can therefore be interpreted as a structured relation between historical lexical profiles and contemporary institutional labels rather than as an artefact of a single random split.

The distinction between flat and hierarchical concordance is equally important. This reading is consistent with the broader scientometric observation that systematic classification failures can encode disciplinary interdependencies (Lyutov et al. 2021). Our setting extends that perspective to a national graduate system with a longitudinally harmonised institutional hierarchy and separates nearby hierarchical divergence from broad cross-domain dispersion. A 49-class Top-1 value near 50% compresses qualitatively different outcomes. Medicine II and Medicine III, for example, have low exact recovery but are predominantly redirected to Medicine I, preserving proximity within the health hierarchy. Materials is rarely recovered exactly but largely remains inside the same CAPES college through Engineering and Chemistry. Cross-college divergence is itself heterogeneous: Teaching is overwhelmingly attracted toward Education/Humanities, Biotechnology toward Life Sciences, whereas Environmental Sciences is distributed across all three colleges. A hierarchical taxonomy combined with destination analysis therefore reveals information that conventional accuracy alone discards.

5.2 Titles, abstracts and keywords as disciplinary signals

Keywords consistently provide the strongest lexical signal under the original pertinence formulation. This pattern is especially relevant because the model is deliberately lexical rather than semantic: it does not infer contextual meaning, syntax or latent semantic relations, but measures how the terms actually chosen by authors are distributed across institutional areas. Author keywords are explicitly selected or created by authors to represent their research, and empirical work has shown that keyword selection draws on document content as well as authors’ prior knowledge and disciplinary background (Lu et al. 2020). Keyword occurrence patterns have also been used as indicators of research specialisation (Sampagnaro 2023), and recent work has explored keyword diversity as a text-based route to interdisciplinarity measurement (Sun and Li 2026). Keywords may therefore occupy an intermediate position between content description and disciplinary self-positioning. The present analysis cannot separate these mechanisms, but this dual role may help explain why keywords align particularly well with a class-aware lexical representation. Titles are likewise highly compressed textual representations. These properties create a close match between the documentary function of T and K and what a frequency-based pertinence profile is able to observe.

This interpretation does not imply that keywords contain more information than abstracts in an absolute sense. Abstracts serve a different communicative function: they express background, objectives, methods, results and conclusions through connected discourse. Consequently, some of their information is carried by contextual and semantic relations that the present representation does not explicitly model. More text therefore need not translate into more discriminative evidence for a purely lexical model. The recurrent ordering observed across the temporal experiments—with K strongest, T intermediate and A weakest—is consequently interpreted as an interaction between field function and model representation, rather than as a general hierarchy of information content. Previous classification research has likewise found

that titles, abstracts and keywords provide distinct signals and should not be treated as interchangeable document fields (Pech et al. 2022).

5.3 Institutional categories and lexical dispersion

The 2019 results also clarify an important conceptual distinction. Institutional classification, lexical proximity and substantive interdisciplinarity are not equivalent. This caution is consistent with research showing that interdisciplinarity indicators depend on their conceptualisation and measurement choices (Zwanenburg et al. 2022; Nakhoda et al. 2023). Recent keyword-based approaches demonstrate that textual diversity can support explicit interdisciplinarity indicators (Sun and Li 2026); the present study adopts a deliberately narrower interpretation. CAPES Evaluation Areas are administrative categories used for graduate evaluation, whereas the present model describes regularities in document vocabulary. A document whose lexical Top-1 area differs from its official area is therefore not necessarily misclassified institutionally; the two labels answer different questions.

The Multidisciplinary Great Area illustrates this distinction particularly well. Teaching, Interdisciplinary, Biotechnology, Materials and Environmental Sciences occupy the same administrative Great Area but exhibit markedly different lexical geometries. Teaching shows a concentrated displacement toward Education/Humanities; Biotechnology is predominantly attracted to Life Sciences; Materials remains largely within the same college; Environmental Sciences spans the three colleges; and Interdisciplinary combines low self-retention with the broadest destination distribution. Administrative co-location within the CAPES hierarchy therefore does not imply a common lexical structure.

Interdisciplinary is noteworthy because it reaches all 49 lexical destinations and has the highest normalised entropy, but this finding should be interpreted narrowly. It indicates that, under the representation investigated here, the institutional Interdisciplinary category does not correspond to a comparably concentrated lexical domain. It does not establish a measure of “true” interdisciplinarity, nor does it imply that CAPES classifications are incorrect. In particular, the Rao–Stirling statistic is used here as a descriptive diversity measure over lexical destinations; uncertainty and validity concerns documented for interdisciplinarity measurement caution against reifying such indices as direct epistemic properties (Nakhoda et al. 2023). The broader epistemic and institutional meaning of this dispersion requires dedicated analysis beyond the scope of the present article.

5.4 Implications for science classification and research evaluation

The analysis suggests a role for transparent lexical models that differs from replacing expert classification. Institutional categories are designed for governance and evaluation, whereas lexical profiles describe regularities in document language. Agreement between the two indicates a concentrated textual signature; systematic nearby divergence indicates overlap between neighbouring institutional domains; broad dispersion

indicates that an administrative category spans multiple lexical regions under the chosen representation.

Such information can complement science maps based on citations, journals or collaboration networks. Theses and dissertations are particularly valuable because they cover graduate research at national scale, including work that may not yet have generated indexed journal publications. At the same time, their institutional labels provide an unusually explicit hierarchical reference system. The CAPES corpus therefore offers a natural laboratory for studying the relation between administrative and textual classifications over long periods.

5.5 Limitations

The method is lexical and frequency-based. It does not model contextual semantics, authorship networks, programme research lines or institutional arrangements. The historical harmonisation reduces naming inconsistencies but cannot eliminate substantive changes in the scope of areas. Results also depend on the completeness and quality of titles, abstracts and keywords across periods. Area size is moderately associated with concordance and dispersion, so area-level contrasts should not be interpreted independently of class support, although similarly sized areas can exhibit sharply different lexical structures. The preprocessing intentionally preserves the original frequency-based design and does not attempt language-specific stemming or semantic normalisation. The study focuses on the original $ME = M_1 + M_2$ formulation; alternative weightings and other text representations are left for future research.

6 Conclusion

This study used more than three decades of Brazilian graduate records to examine whether institutional disciplinary boundaries are recoverable from transparent lexical profiles and how divergences should be interpreted when they are not. The method shows comparatively consistent out-of-time behaviour, with keywords providing the strongest representation. In 2019, 49.69% of 94,194 keyword-bearing theses and dissertations were assigned to their official Evaluation Area, 63.51% remained within the same Great Area or closer, and 79.18% remained within the same CAPES college or closer.

The principal contribution is structural rather than predictive. By combining lexical rankings with the CAPES hierarchy, the study distinguishes exact concordance, nearby institutional divergence, concentrated cross-domain displacement and broad lexical dispersion. The 2019 results show that these are not interchangeable phenomena: Teaching is strongly attracted toward Education/Humanities, Biotechnology toward Life Sciences, Materials largely remains within its college, and Environmental Sciences spans all three colleges. Interdisciplinary provides the broadest destination distribution, while Teaching demonstrates that large hierarchical displacement can coexist with low entropy.

These findings support a view of national research taxonomies as interfaces between administrative organisation and the evolving textual structure of knowledge. Lexical profiles do not identify a “true” disciplinary home and should not be used

to audit individual programmes. They can, however, provide an interpretable layer for examining where institutional boundaries are lexically sharp, where neighbouring fields overlap, and where categories span multiple regions of the scientific vocabulary. Future work may compare these lexical structures with citation, collaboration and semantic-embedding maps while preserving the longitudinal design used here.

Supplementary information. Additional audited results are provided in Online Resource 1, including the harmonisation map, temporal-validation and $M_1/M_2/ME$ ablation results, and the 49-area hierarchical results for the 2019 keyword application.

Acknowledgements. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Declarations

Funding This study was financed in part by CAPES - Finance Code 001.

Competing interests The authors declare that they have no competing interests.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Data availability The source data are public and available from the CAPES Open Data Portal. The article reports aggregate tables derived from those public records and from the harmonisation procedure described in the manuscript.

Code availability Custom scripts were developed to process the public CAPES records and reproduce the analyses presented. The source code is not publicly available in this version. The formulas, temporal protocol, computational procedures and aggregate results required to evaluate the method are described in the manuscript and Supplementary Information.

Author contributions Emerson Cordeiro Silva: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft. Paulo Eduardo Ambrosio: Formal analysis, Writing – review & editing. Francisco Bruno Souza Oliveira: Conceptualization, Methodology, Software, Formal analysis, Validation, Visualization, Supervision, Writing – review & editing, Project administration. All authors read and approved the final manuscript.

ORCID Paulo Eduardo Ambrosio: <https://orcid.org/0000-0002-6043-3591>; Francisco Bruno Souza Oliveira: <https://orcid.org/0000-0002-9847-7290>.

References

Attieh, J., Tekli, J.: Supervised term-category feature weighting for improved text classification. *Knowledge-Based Systems* **261**, 110215 (2023) <https://doi.org/10.1016/j.knosys.2022.110215>

CAPES: Sobre as áreas de avaliação. <https://www.gov.br/capes/pt-br/aceso-a-informacao/acoes-e-programas/avaliacao/sobre-a-avaliacao/areas-avaliacao/sobre-as-areas-de-avaliacao/sobre-as-areas-de-avaliacao/>.

Accessed 28 Aug 2026 (2024)

- CAPES: Tabela de Areas de Conhecimento/Avaliacao. <https://www.gov.br/capes/pt-br/aceso-a-informacao/acoes-e-programas/avaliacao/instrumentos/documentos-de-apoio/tabela-de-areas-de-conhecimento-avaliacao>. Accessed 28 Aug 2026 (2024)
- CAPES: 1987 a 2012 - Catalogo de Teses e Dissertacoes - Brasil. <https://dadosabertos.capes.gov.br/dataset/1987-a-2012-catalogo-de-teses-e-dissertacoes-brasil>. Accessed 28 Aug 2026 (2026)
- CAPES: 2017 a 2020 - Catalogo de Teses e Dissertacoes - Brasil. <https://dadosabertos.capes.gov.br/dataset/2017-2020-catalogo-de-teses-e-dissertacoes-da-capes>. Accessed 28 Aug 2026 (2026)
- CAPES: Catalogo de Teses e Dissertacoes de 2013 a 2016. <https://dadosabertos.capes.gov.br/dataset/catalogo-de-teses-e-dissertacoes-de-2013-a-2016>. Accessed 28 Aug 2026 (2026)
- Chen, K., Zhang, Z., Long, J., Zhang, H.: Turning from tf-idf to tf-igm for term weighting in text classification. *Expert Systems with Applications* **66**, 245–260 (2016) <https://doi.org/10.1016/j.eswa.2016.09.009>
- Debole, F., Sebastiani, F.: Supervised term weighting for automated text categorization. In: *Proceedings of the 2003 ACM Symposium on Applied Computing*, pp. 784–788 (2003). <https://doi.org/10.1145/952686.952688>
- Dogan, T., Uysal, A.K.: A novel term weighting scheme for text classification: Tf-mono. *Journal of Informetrics* **14**(4), 101076 (2020) <https://doi.org/10.1016/j.joi.2020.101076>
- Goh, Y.C., Cai, X.Q., Theseira, W., Ko, G., Khor, K.A.: Evaluating human versus machine learning performance in classifying research abstracts. *Scientometrics* **125**, 1197–1212 (2020) <https://doi.org/10.1007/s11192-020-03614-2>
- Guns, R., Sile, L., Eykens, J., Verleysen, F.T., Engels, T.C.E.: A comparison of cognitive and organizational classification of publications in the social sciences and humanities. *Scientometrics* **116**(2), 1093–1111 (2018) <https://doi.org/10.1007/s11192-018-2775-x>
- Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A survey on concept drift adaptation. *ACM Computing Surveys* **46**(4), 1–37 (2014) <https://doi.org/10.1145/2523813> . Article 44
- Lu, W., Liu, Z., Huang, Y., Bu, Y., Li, X., Cheng, Q.: How do authors select keywords? a preliminary study of author keyword selection behavior. *Journal of Informetrics* **14**(4), 101066 (2020) <https://doi.org/10.1016/j.joi.2020.101066>

- Lan, M., Tan, C.L., Su, J., Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(4), 721–735 (2009) <https://doi.org/10.1109/TPAMI.2008.110>
- Lyutov, A., Uygun, Y., Hütt, M.-T.: Machine learning misclassification of academic publications reveals non-trivial interdependencies of scientific disciplines. *Scientometrics* **126**, 1173–1186 (2021) <https://doi.org/10.1007/s11192-020-03789-8>
- Leydesdorff, L., Wagner, C.S., Bornmann, L.: Interdisciplinarity as diversity in citation patterns among journals: Rao–stirling diversity, relative variety, and the gini coefficient. *Journal of Informetrics* **13**(1), 255–269 (2019) <https://doi.org/10.1016/j.joi.2018.12.006>
- McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018) <https://doi.org/10.48550/arXiv.1802.03426>
- Mutz, R.: Diversity and interdisciplinarity: Should variety, balance and disparity be combined as a product or better as a sum? an information-theoretical and statistical estimation approach. *Scientometrics* **127**, 7397–7414 (2022) <https://doi.org/10.1007/s11192-022-04336-3>
- Nakhoda, M., Whigham, P., Zwanenburg, S.: Quantifying and addressing uncertainty in the measurement of interdisciplinarity. *Scientometrics* **128**, 6107–6127 (2023) <https://doi.org/10.1007/s11192-023-04822-2>
- Pech, G., Delgado, C., Sorella, S.P.: Classifying papers into subfields using abstracts, titles, keywords and keywords plus through pattern detection and optimization procedures: an application in physics. *Journal of the Association for Information Science and Technology* **73**(11), 1513–1528 (2022) <https://doi.org/10.1002/asi.24655>
- Sampagnaro, G.: Keyword occurrences and journal specialization. *Scientometrics* **128**, 5629–5645 (2023) <https://doi.org/10.1007/s11192-023-04815-1>
- Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Information Processing & Management* **24**(5), 513–523 (1988) [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)
- Shannon, C.E.: A mathematical theory of communication. *Bell System Technical Journal* **27**(3), 379–423 (1948) <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Silva, E.C.: Modelagem para classificacao e pertinencia de produtos academicos dos programas credenciados pela capes. Master’s thesis, Universidade Estadual de Santa Cruz, Ilheus (2022)

- Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. *Information Processing & Management* **45**(4), 427–437 (2009) <https://doi.org/10.1016/j.ipm.2009.03.002>
- Sun, J., Li, Y.: Interdisciplinarity in academic research: a keyword-based approach to knowledge diversity. *Scientometrics* **131**(5), 3679–3698 (2026) <https://doi.org/10.1007/s11192-026-05651-9>
- Shu, F., Ma, Y., Qiu, J., Larivière, V.: Classifications of science and their effects on bibliometric evaluations. *Scientometrics* **125**(3), 2727–2744 (2020) <https://doi.org/10.1007/s11192-020-03701-4>
- Stirling, A.: A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface* **4**(15), 707–719 (2007) <https://doi.org/10.1098/rsif.2007.0213>
- Zwanenburg, S., Nakhoda, M., Whigham, P.: Toward greater consistency and validity in measuring interdisciplinarity: a systematic and conceptual evaluation. *Scientometrics* **127**, 7769–7788 (2022) <https://doi.org/10.1007/s11192-022-04310-z>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1.xlsx](#)