

Beyond Embeddings: Comparing Language-Aware Sparse Representations and Fine-Tuned Multilingual Transformer Classifiers for Multi-label Subject Classification in Institutional Repositories

Carlos Javier Nusch^{1,2}, Leticia Cecilia Cagnina^{3,4}, Ariel Lira^{1,2}, Gonzalo Villarreal^{1,2}, Leandro Antonelli^{5,6}, Marcelo Luis Errecalde³, Rafael Bértoli^{1,2}, Manuel Rojas^{1,2}, María Marta Vila^{1,2}, and Marisa De Giusti^{1,2}

¹ Dirección PREBI-SEDICI, Universidad Nacional de La Plata, Argentina

² CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina

³ LIDIC, Facultad de Ciencias Físico Matemáticas y Naturales, Universidad Nacional de San Luis, Argentina

⁴ Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina

⁵ LIFIA, Facultad de Informática, Universidad Nacional de La Plata, Argentina

⁶ CAETI, Facultad de Tecnología Informática, Universidad Abierta Interamericana, Argentina

{carlosnusch,gonzalo}@prebi.unlp.edu.ar;

{lcagnina,merreca}@email.unsl.edu.ar

{alira,bertoli.rafa,rojasmanuel,vilamm,marisa.degiusti}@sedici.unlp.edu.ar;

lanto@lifia.info.unlp.edu.ar

Corresponding author: carlosnusch@prebi.unlp.edu.ar

Abstract. Institutional repositories need scalable support for multilingual subject metadata assignment to long documents. On 19,710 SEDICI items and 37 labels, we compare 231 sparse/fixed-embedding configurations and two fine-tuned DistilUSE/LaBSE classifiers under separate threshold-selection, validation, and test partitions. Language-aware BM25 full text with LinearSVC achieved held-out Macro-F1 = 0.7522 (95% CI 0.7372–0.7643), Micro-F1 = 0.7685 (0.7567–0.7797), and Recall@3 = 0.9320; the LaBSE-initialized classifier reached Macro-F1 = 0.7069. A post-hoc BM25–LaBSE matched-exposure control, validation-only over three seeds with no test access, preserved the aggregate BM25 advantage. Strong sparse baselines therefore remain competitive under the evaluated Transformer protocols.

Keywords: multi-label classification · automated subject indexing · institutional repositories · multilingual preprocessing · BM25 · DistilUSE · LaBSE

1 Introduction

Institutional repositories contain large and heterogeneous collections whose metadata must remain sufficiently consistent to support browsing, discovery, and reuse. Subject metadata is especially demanding because assignment normally requires intellectual analysis by repository staff. Automatic classification is therefore most useful as decision support: a system proposes candidate subjects that catalogers accept, reject, or modify rather than replacing professional judgment [1,2]. The task is naturally multi-label because a single scholarly item may legitimately receive several subjects [3].

SEDICI, the institutional repository of Universidad Nacional de La Plata, provides a realistic multilingual, long-document setting with multiple abstracts or full-text files. Earlier SEDICI work found TF-IDF with linear classifiers highly competitive on abstracts and keywords [4]; a later full-text/BM25 experiment used less comprehensive Transformer baselines. Here we add language-aware processing, larger full-text budgets, hierarchical fixed embeddings, supervised Transformer fine-tuning, separate decision-threshold selection, and uncertainty analysis.

We ask whether fine-tuned multilingual Transformers outperform strong sparse baselines under this protocol. Contributions are: (1) a duplicate-aware reproducible corpus workflow preserving text units; (2) audited language-aware pre-processing; (3) 231 sparse/fixed-embedding configurations plus DistilUSE/LaBSE-initialized classifiers; (4) train-threshold-selection-validation-test evaluation with bootstrap uncertainty and a convergence audit; and (5) a post-hoc BM25–LaBSE matched-exposure control evaluated only on validation, with the test partition untouched.

2 Related Work

Automatic subject assignment combines text classification with indexing and knowledge organization. Sebastiani systematized machine-learning approaches to text categorization [5], while multi-label learning formalizes cases where several labels apply simultaneously [3]. In libraries, automated suggestions are also evaluated as aids to retrieval and professional indexing [1]; Annif exemplifies this cataloger-support perspective [2], while Ibero-American work has explored controlled-vocabulary and lexical approaches [6].

Sparse and dense representations encode different assumptions. TF-IDF and BM25 emphasize lexical evidence; BM25 additionally models term-frequency saturation and document-length normalization [7,8]. Multilingual sentence-embedding models such as DistilUSE [10]¹ and LaBSE [9] instead produce dense sentence representations. Long-document use is non-trivial because transformer encoders

¹ DistilUSE model: <https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1>; LaBSE model: <https://huggingface.co/sentence-transformers/LaBSE>.

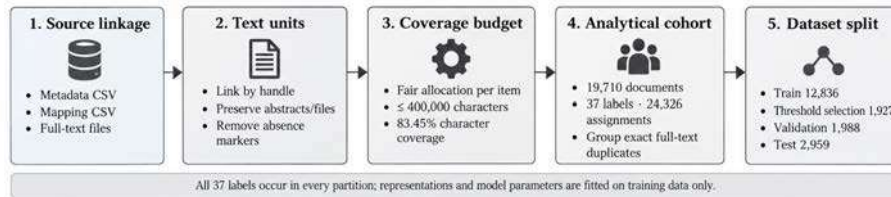


Fig. 1. Dataset construction and stratified partitioning workflow.

have limited input windows, motivating segmentation and hierarchical aggregation strategies [11]. In multi-label settings, threshold choice also matters because a single default cutoff can be suboptimal across heterogeneous labels [12].

Within SEDICI, prior work selected 37 frequent labels and compared lexical and multilingual representations over abstracts and keywords [4]. Separate studies evaluated language identification and model adaptation for repository abstracts, documenting metadata inconsistencies and the value of language-aware curation [13,14]. Here, current language identification drives preprocessing and the earlier audited reference evaluates the language component.

3 Data and Corpus Construction

3.1 Sources, Linking, and Text Units

The experiment uses an item-level metadata export from SEDICI, a mapping between public handles and internal DSpace Assetstore identifiers, and extracted full-text files. Assetstore identifiers are linked back to handles, which serve as the common key for metadata, abstracts, keywords, targets, and full text. The 37-label subset follows the coverage analysis established in the previous SEDICI subject-classification study [4].

Text is preserved at unit level: DSpace-delimited abstracts remain independent segments and multiple full-text files per handle remain separate units; explicit absence markers are removed. A fair handle-level budget prevents the first file from consuming the allowance, capped at 400,000 full-text characters per item. After linking and unit construction, the consolidated corpus and analytical dataset were materialized as compressed Parquet files to avoid repeated ingestion and full-text processing. The corpus contains 19,710 documents, 24,326 assignments, 24,411 abstract segments, and 20,750 full-text files; 4,365 documents have multiple abstracts and 463 multiple full texts. Fig. 1 summarizes the workflow.

3.2 Coverage and Partitioning

Across the analytical cohort, the source full texts contain 1.213 billion characters, of which 1.012 billion are represented under the fair per-item budget, corresponding to 83.45% character coverage. Only 510 documents (2.59%) reach

Table 1. Analytical corpus and partition statistics.

Property	Value
Documents	19,710
Subject labels	37
Label assignments	24,326
Abstract segments	24,411
Full-text files	20,750
Full-text character coverage	83.45%
Documents reaching 400k cap 510	(2.59%)
Train	12,836
Threshold selection	1,927
Validation	1,988
Test	2,959
Labels present in every split	37 / 37

the 400,000-character cap. Full texts that are identical after Unicode-space normalization and case folding are assigned to the same content group; iterative multi-label stratification is then applied to these groups [15], preventing equivalent represented content from crossing partitions. The split contains 12,836 training, 1,927 threshold-selection (pipeline name: calibration), 1,988 validation, and 2,959 test items; all 37 labels occur throughout. Training fits parameters, threshold selection chooses cutoffs, validation selects configurations/epochs, and test is held out. No probability/score calibration is performed (Table 1).

4 Multilingual Text Processing

Language identification is applied independently to abstract and full-text units with `langid.py` [16]. Cataloged language values are retained as evidence but not assumed to match the text; long or multilingual units are also inspected through windows. Current LangID drives preprocessing, while historical language evidence is used only for evaluation and never supplies subject labels.

Three sparse variants are evaluated: raw (custom preprocessing bypassed, while vectorizers lowercase internally); normalized (Unicode NFKC, null-to-space replacement, whitespace collapse, and trimming); and language-aware, which additionally tokenizes words and removes language-specific stopwords obtained through the `stopwordsiso` 0.7.1 Python package, which distributes the Stopwords ISO collection [17]. Keywords retain all terms. For `mul`, `und`, or unsupported languages, stopword removal is skipped. No stemming or lemmatization is used. Transformer inputs receive the same normalization but no stopword removal.

Declared/detected agreement is 98.51% for 15,166 auditable abstract records and 96.05% for 19,675 full-text records. Abstract LangID is also evaluated against a historical audited reference of 23,609 matched segments: 23,055 come from human/historical-LangID agreement and 554 disagreements were manually audited. Accuracy is 99.87%; these cases are therefore not all independent manual annotations. Table 2 summarizes the audit and Fig. 2 the protocol.

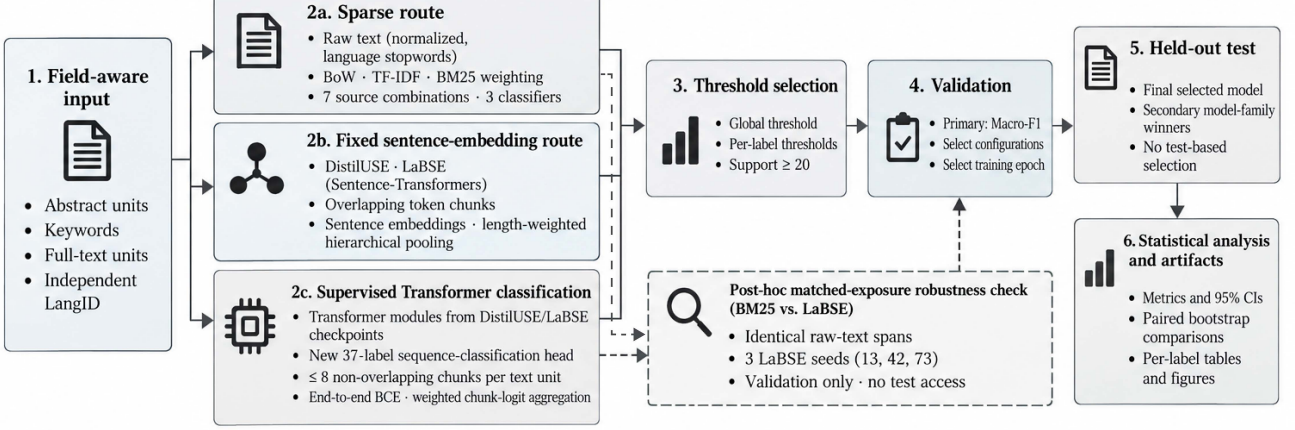


Fig. 2. End-to-end experimental workflow.

Table 2. Language-identification audit used to validate language-aware preprocessing.

Evaluation	N Agreement / accuracy	
Declared vs detected: abstract	15,166	98.51%
Declared vs detected: full text	19,675	96.05%
Historical audited abstract reference	23,609 segments	99.87%

5 Experimental Design

5.1 Text Sources and Representations

Seven source combinations are evaluated: abstract, keywords, full text, and their four combinations. Training-only sparse representations are BoW, TF-IDF, and BM25 with word uni/bigrams, up to 100,000 features, `min_df`=2, and `max_df`=0.98. BM25 uses $k_1 = 1.5$, $b = 0.75$, and the positive $\text{IDF}(t) = \log\left(1 + \frac{N - \text{df}(t) + 0.5}{\text{df}(t) + 0.5}\right)$; vocabulary and collection statistics are fitted on training only, and the resulting BM25-weighted sparse document-term features are supplied to the linear classifiers [7].

Fixed sentence-embedding baselines use multilingual DistilUSE and LaBSE with unchanged parameters. Long inputs use overlapping chunks (32-token overlap; maximum 128 tokens for DistilUSE and 256 for LaBSE); embeddings are length-weighted and pooled hierarchically from chunk to text unit to handle. All-field weights are 0.25 abstract, 0.15 keywords, and 0.60 full text.

The supervised Transformer path is distinct: Hugging Face sequence classifiers are initialized from the Transformer module of each DistilUSE/LaBSE checkpoint, given a new 37-output multi-label head, and fine-tuned end-to-end (backbone + head), rather than adapting the complete Sentence-Transformers pipeline with a similarity/contrastive objective. Up to eight uniformly sampled non-overlapping 256-token chunks per text unit contribute document-weighted logits. Training uses `BCEWithLogitsLoss` with $N_{\text{negative}}/N_{\text{positive}}$ positive-class weights, AdamW (2×10^{-5}), batch 16, evaluation batch 64, gradient accumulation 4, gradient checkpointing, and early stopping (patience 2; minimum improvement 0.002),

restoring the best checkpoint. Both variants selected epoch 5; DistilUSE improved monotonically and LaBSE had a small epoch-4 fluctuation.

5.2 Classifiers, Threshold Selection, and Model Selection

Sparse and fixed sentence-embedding routes use One-vs-Rest Logistic Regression, LinearSVC, and SGD with logistic loss [18,19,20]. The grid has 189 sparse + 42 fixed-embedding configurations; two supervised Transformer classifiers bring the total to 233 validation candidates. A later eight-fit LinearSVC convergence audit does not reopen the broader search.

The 1,927-item threshold-selection partition tunes per-label decision cutoffs when support is at least 20; lower-support labels inherit the global cutoff. This is threshold optimization, not probability calibration. Validation Macro-F1 is primary. Pre-specified family winners receive secondary held-out evaluation but cannot replace the global validation-selected configuration.

The initial winner was `language_stopwords` + full text + BM25 + LinearSVC. A Liblinear warning triggered a focused audit over $C \in \{1.0, 0.5\}$ and `tol` $\in \{0.0001, 0.0005\}$ on the identical cohort and split. $C = 1.0$ failed only for Humanidades; both $C = 0.5$ variants converged for all 37 labels and tied on every stored validation metric. The implementation returned `tol` = 0.0005 first, so it was retained; no superiority over `tol` = 0.0001 is claimed. All 37 one-vs-rest fits converged again in the final refit. The audit was prompted by optimizer behavior after the initial held-out evaluation, and test performance did not choose C or `tol`.

Post-hoc matched-exposure control. On the frozen 16,751 non-test items, a validation-only control restricted BM25 to exact raw spans selected by the LaBSE tokenizer: at most eight non-overlapping 256-token sequences (254 content tokens plus special tokens) per full-text unit (15.78% of capped source characters; multi-file items could exceed eight overall). BM25 received no other lexical content; frozen upstream LangID only selected stopword lists. LaBSE was fine-tuned end-to-end for up to 20 epochs (seeds 13/42/73); thresholds came only from threshold selection and checkpoints from validation Macro-F1 (patience 2; minimum improvement 0.002). Test ($n = 2,959$) remained untouched.

5.3 Metrics and Uncertainty

Metrics include subset accuracy, Hamming loss, micro/macro/sample precision, recall and F1, average precision, and precision/recall at $k = 1, 3, 5$. Macro-F1 is primary. The final model receives 1,000-resample bootstrap intervals; paired family differences use the same 2,959 test handles and seed 42 against stored family-winner predictions from the main v3 run. Here, “v3” denotes the third version of that experimental run.

Reproducibility. Main run: 20260802T165746Z (7181fcceb7a6); convergence audit: 20260804T102544Z (6e46e5c2af02). Main environment: Python 3.12.13,

Table 3. Pre-specified model-family winners and final held-out test estimates.

Family / selected configuration	Val. Macro-F1	Held-out Macro-F1	Held-out Micro-F1
BM25 / FT / LinearSVC ($C = 0.5$)	0.7497	0.7522	0.7685
TF-IDF / A+FT / LinearSVC	0.7357	0.7338	0.7529
LaBSE-init. seq. classifier / A+K+FT	0.7148	0.7069	0.7279
BoW / A+K+FT / Logistic Reg.	0.6895	0.6782	0.7135
DistilUSE-init. seq. classifier / A+K+FT	0.6663	0.6595	0.6746
Fixed DistilUSE emb. / K+FT / LinearSVC	0.6278	0.6211	0.6397
Fixed LaBSE emb. / K+FT / LinearSVC	0.6254	0.6143	0.6334

Note. A = abstract; K = keywords; FT = full text; init. = sequence classifier initialized from the checkpoint Transformer module; emb. = fixed sentence embeddings. Family-level test estimates are secondary analyses; the BM25 row is the global validation-selected configuration refined only by the focused convergence audit.

scikit-learn 1.6.1, PyTorch 2.11.0, Transformers 5.13.1, Sentence Transformers 5.6.0, stopwordsiso 0.7.1. Matched-exposure artifacts record pinned code/model revisions. Supplementary artifacts are public.²

6 Results

6.1 Global Selection and Representation Families

In the v3 validation grid, language-aware full-text BM25 + LinearSVC produced the best Macro-F1 (0.7483), whereas BM25 over all three fields + Logistic Regression produced the best Micro-F1 (0.7774; Macro-F1 = 0.7388). Because Macro-F1 was prespecified as primary, the full-text LinearSVC model advanced to the convergence audit. With $C = 0.5$, both tolerances converged for all 37 one-vs-rest fits and tied on stored validation metrics; retained `tol = 0.0005` reached validation Macro-F1 = 0.7497 and Micro-F1 = 0.7732. Final held-out Macro-F1 and Micro-F1 were 0.7522 and 0.7685. Table 3 compares the selected model-family winners.

Relative to the final BM25-weighted LinearSVC model, paired Macro-F1 differences are -0.0184 for TF-IDF (95% CI -0.0299 to -0.0068), -0.0453 for the LaBSE-initialized classifier (-0.0587 to -0.0311), and -0.0928 for the DistilUSE-initialized classifier (-0.1070 to -0.0779). All remaining family intervals are also below zero; no bootstrap replicate favored TF-IDF. These secondary analyses support, rather than determine, the model-family-level conclusion.

6.2 Final Held-Out Performance and Cataloger-Oriented Ranking

Table 4 reports the final held-out performance and cataloger-oriented ranking metrics.

Recall@3 = 0.9320 is particularly relevant to a recommendation workflow: when three candidate subjects are shown, the ranked suggestions recover a high proportion of the true assignments, although precision decreases as more candidates are displayed. This supports evaluating the classifier as a human-in-the-loop suggestion system rather than only through exact-set prediction.

² Repository: <https://github.com/KharolusIII/institutional-repository-subject-classification>.
Reproducibility: https://github.com/KharolusIII/institutional-repository-subject-classification/tree/f8520ae/paper_artifacts/reviewer_response/exposure_parity.

Table 4. Final held-out test performance of the convergence-resolved BM25-weighted LinearSVC model.

Metric	Value	Metric	Value
Subset accuracy	0.5863	Hamming loss	0.0164
Micro precision / recall	0.7252 / 0.8172	Micro-F1 (95% CI)	0.7685 (0.7567–0.7797)
Macro precision / recall	0.7409 / 0.7899	Macro-F1 (95% CI)	0.7522 (0.7372–0.7643)
Sample-F1	0.7631	Precision@1 / Recall@1	0.8280 / 0.7262
Precision@3 / Recall@3	0.3810 / 0.9320	Precision@5 / Recall@5	0.2377 / 0.9660

Table 5. Matched-exposure validation. $\Delta = \text{BM25} - \text{LaBSE}$, computed from unrounded values; no test evaluation.

Model/Seed	Epoch	Macro-F1	Micro-F1	Δ Macro
LaBSE 13	14	0.7419	0.7551	0.0154
LaBSE 42	9	0.7395	0.7518	0.0178
LaBSE 73	9	0.7380	0.7542	0.0194
LaBSE mean	–	0.7398	0.7537	0.0175
BM25 refit	–	0.7573	0.7701	–

6.3 Supervised Transformer Sequence Classification and Subject-Level Behavior

The LaBSE-initialized classifier reaches held-out Macro-F1 = 0.7069 versus 0.6143 for fixed LaBSE embeddings (+0.0927); DistilUSE reaches 0.6595 versus 0.6211 (+0.0384). These descriptive family comparisons are not controlled ablations because architecture and training objective differ. Both supervised variants peak at validation epoch 5: DistilUSE rises from 0.5967 to 0.6663, and LaBSE from 0.6596 to 0.7148 with a small epoch-4 dip.

Subject-level performance remains heterogeneous. Odontología (F1 = 0.9481), Ciencias Informáticas (0.9420), and Educación Física (0.9396) are among the strongest labels; Biología (0.3009), Ciencias Sociales (0.4590), and Urbanismo (0.5542) are among the hardest. Humanidades has support 227 but F1 = 0.6571, so frequency alone does not explain difficulty.

Matched-exposure robustness. On validation (Table 5), BM25 Macro/Micro-F1 was 0.757330/0.770086. LaBSE epochs 14/9/9 (seeds 13/42/73) averaged 0.739794 ± 0.001979 / 0.753722 ± 0.001700 (sample SD); mean BM25–LaBSE gaps were 0.017535/0.016364 and BM25 led all runs. No test inference/evaluation or SBERT/DistilUSE extension was performed.

7 Discussion

The research question receives a qualified negative answer: with separate decision-threshold selection, neither DistilUSE- nor LaBSE-initialized classifier surpassed the strongest sparse configuration by prespecified Macro-F1. Family differences remain descriptive because architectures and chunk handling differ.

BM25 is well matched to terminology-rich repository text: language-aware full text maximizes validation Macro-F1, while all fields with Logistic Regression maximize Micro-F1. More importantly, BM25 retained the aggregate validation

advantage over LaBSE in all three matched-exposure seeds. Unequal content exposure therefore does not alone explain the ordering, although tokenization, preprocessing, representation, aggregation, and inductive biases remain different; subject-level behavior is also heterogeneous.

LaBSE fine-tuning nonetheless closes much of the fixed-embedding gap. Recall@3 = 0.9320 supports assistive ranking; prospective work should measure cataloger acceptance, correction effort, threshold stability, and—if calibrated confidence estimates are introduced—calibration drift.

8 Limitations and Future Work

The study covers one repository and 37 frequent labels. Supervised classification samples at most eight chunks per text unit; the post-hoc matched-exposure control addresses this imbalance only for BM25 versus LaBSE on the frozen validation split. Its three seeds estimate within-split training variability, not split-level variability, and the original family comparison remains single-run. The language benchmark is a hybrid historical abstract reference.

The convergence audit followed an optimization warning after an earlier test read, a procedural limitation despite validation-based C/tol choices. We optimized decision thresholds but did not calibrate probabilities/scores. Future work should add repeated/external/temporal partitions, broader multiseed optimization, hierarchical long-document Transformers with learned aggregation, explicit probability/score calibration, long-tail evaluation, and prospective cataloger studies.

9 Conclusion

On 19,710 SEDICI items and 37 labels, language-aware BM25 full text with LinearSVC achieved the highest held-out Macro-F1 among the evaluated configurations (0.7522); LaBSE was the stronger of the two evaluated supervised Transformer classifiers (Macro-F1 = 0.7069). BM25’s aggregate validation advantage also persisted across three LaBSE seeds under matched raw-text exposure; no test inference or evaluation was performed for that post-hoc control. These bounded comparisons do not establish intrinsic sparse-model superiority. Recall@3 = 0.9320 supports assistive recommendation; broader validation, hierarchical Transformers, explicit probability/score calibration, and prospective cataloger studies remain future work.

Author Contributions. Following TaDiRAH, Nusch: Programming, Modeling, Encoding, Analyzing, Cleaning, Evaluating, Enriching, Writing; Cagnina, Lira, Villarreal, Antonelli, Errecalde, De Giusti: Reviewing, Editing; Bértoli, Rojas, Vila: Extracting, Collecting, Preparing DSpace data.

Disclosure of Interests. The authors declare no competing interests.

References

1. Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M., Hiom, D.: A framework for evaluating automatic indexing or classification in the context of retrieval. *J. Assoc. Inf. Sci. Technol.* **67**, 3–16 (2016).
2. Suominen, O.: Annif: DIY automated subject indexing using multiple algorithms. *LIBER Q.* **29**, 1–25 (2019).
3. Zhang, M.-L., Zhou, Z.-H.: A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **26**, 1819–1837 (2014).
4. Nusch, C.J., Cagnina, L.C., Peloché, S.B., Villarreal, G.L., Lira, A.J., Antonelli, R.L., Folegatto, L.E., Errecalde, M.L., De Giusti, M.R.: Clasificación automática de materias en repositorios institucionales mediante aprendizaje supervisado y representaciones vectoriales multilingües: un estudio de caso en SEDICI. In: XIV BIREDIAL-ISTEC (2025)
5. Sebastiani, F.: Machine learning in automated text categorization. *ACM Comput. Surv.* **34**, 1–47 (2002).
6. González-Pérez, C., Liy, J.I.V., García, A.G., Ibáñez, P.C.: Clasificación temática automática de documentos basada en vocabularios y frecuencias de uso. El caso de artículos de divulgación científica. *Rev. Esp. Doc. Científica* **46**, e362 (2023).
7. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **3**, 333–389 (2009).
8. Salton, G., McGill, M.J.: *Introduction to Modern Information Retrieval*. McGraw-Hill, New York (1983)
9. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT Sentence Embedding. In: *Proc. ACL 2022*, pp. 878–891. ACL (2022).
10. Reimers, N., Gurevych, I.: Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In: *Proc. EMNLP 2020*, pp. 4512–4525. ACL (2020).
11. Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., Dehak, N.: Hierarchical Transformers for Long Document Classification. In: *Proc. ASRU 2019*, pp. 838–844 (2019).
12. Lipton, Z.C., Elkan, C., Naryanaswamy, B.: Optimal Thresholding of Classifiers to Maximize F1 Measure. In: *ECML PKDD 2014*, pp. 225–239. Springer (2014).
13. Nusch, C.J., Cagnina, L.C., Errecalde, M.L., Antonelli, L., De Giusti, M.R.: Detección de idiomas como tarea de curaduría de datos para repositorios institucionales: desempeño de bibliotecas disponibles y modelos de lenguaje. In: *BIREDIAL-ISTEC 2024* (2025)
14. Nusch, C.J., Cagnina, L.C., Lira, A.J., Villarreal, G.L., Antonelli, R.L., Errecalde, M.L., De Giusti, M.R., Tettamanti, S.: Detección automática de idiomas en textos cortos de repositorios institucionales: ¿vale la pena realizar un ajuste fino sobre un modelo de lenguaje? In: XIV BIREDIAL-ISTEC (2025)
15. Sechidis, K., Tsoumakas, G., Vlahavas, I.: On the Stratification of Multi-label Data. In: *ECML PKDD 2011*, pp. 145–158. Springer (2011).
16. Lui, M., Baldwin, T.: *langid.py*: An Off-the-shelf Language Identification Tool. In: *Proc. ACL 2012 System Demonstrations*, pp. 25–30. ACL (2012)
17. Suriyawongkul, A., Díaz, G.: *stopwordsiso 0.7.1*: Python interface to the Stopwords ISO multilingual collection. PyPI (2026)
18. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **9**, 1871–1874 (2008)
19. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
20. Bottou, L.: Large-Scale Machine Learning with Stochastic Gradient Descent. In: *COMPSTAT 2010*, pp. 177–186. Physica-Verlag (2010)