Benchmarking Text Classifiers for Economic News in Brazilian Portuguese: In-Domain Quality and Out-of-Distribution Generalization

Diego de Almeida Miranda[1], **João Assaoka**[1] **and Lilian Berton**[1]

[1]Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brasil

{almeida.diego,joao.assaoka,lberton}@unifesp.br

Abstract. *Filtering economic news is an upstream triage step for finance-oriented NLP, yet for Brazilian Portuguese there is no standardised comparison of the available model families. We benchmark TF–IDF linear models, Portuguese BERT encoders (BERTimbau, FinBERT-PT-BR, NorBERTo, DeB3RTa), and instruction-tuned LLMs on a corpus of* Folha de São Paulo *articles, under two shared-split formulations: a binary* `mercado`*-vs-rest task and a* $7+$`other` *multiclass task over editorial sections. Every system is evaluated on identical 5-fold splits with a calibration- and significance-aware protocol (Brier, ECE, Bonferroni-corrected McNemar) and a cross-validation-only selection funnel that never consults the test set. A stacking ensemble is the strongest in-domain filter (multiclass macro-F1* 0.94*) and BERT leads among individual models, but a TF–IDF model recovers most of that performance at one-to-two orders of magnitude lower cost and stays competitive on a zero-shot out-of-domain probe over four external corpora — the trade-off that matters for an upstream triage stage.*

1. Introduction

Misinformation and information asymmetry within the economic and financial domains exert measurable, sometimes devastating, effects on market stability, asset pricing, and retail investors [Araci 2019, Yang et al. 2020]. Recent Brazilian incidents — such as fabricated statements falsely attributed to the Central Bank president that immediately triggered exchange rate volatility, alongside coordinated digital campaigns aimed at discrediting official macroeconomic statistics — render this phenomenon deeply concrete. Because manual fact-checking and expert validation cannot scale to match the velocity of modern news streams, automated natural language processing (NLP) pipelines have become vital [Monteiro et al. 2018, Garcia et al. 2022]. Following the established tradition of check-worthy claim detection, these pipelines deploy cheap *upstream* triage stages ahead of resource-intensive downstream verification layers.

A foundational component of such pipelines is a robust *domain filter* designed to forward only economically-relevant textual material downstream. In this specific operational context, a false negative (missing a critical economic report) is strictly costlier than a false positive (filtering a generic article). This structural asymmetry mandates a recall-first architectural design and motivates the use of F_1-score (or macro-F_1), rather

than raw accuracy, as the primary optimization metric under the corpus's natural class imbalance ($\pi \approx 12.5\%$). While the Portuguese NLP ecosystem has seen a rapid proliferation of sparse baselines, deep contextual encoders like BERTimbau [Souza et al. 2020] or FinBERT-PT-BR [Santos et al. 2023], and instruction-tuned Large Language Models (LLMs) [Pires et al. 2023, Garcia et al. 2024a, Corrêa et al. 2025], these distinct model families have not, to our knowledge, been systematically benchmarked on the *same* splits, under a unified tuning budget, and via a strict calibration- and significance-aware protocol for this task.

We address this gap by studying two parallel machine learning formulations over a comprehensive corpus of *Folha de São Paulo* articles. The binary task maps an article directly to $\{\texttt{mercado}, \texttt{outros}\}$, strictly preserving the $\approx 12.5\%$ positive prevalence across every cross-validation fold and the held-out test set. Conversely, the multiclass task maps text to the top-7 editorial sections plus a generic `outros` class, explicitly situating the economic domain against its structural editorial competitors. Crucially, both formulations share identical document split indices by construction — only the target label vector changes — ensuring that performance variations between tasks are not confounded with dataset sampling noise (§3.1). A cross-validation-based funnel (§3) advances the best-performing individual models and ensemble configurations to an out-of-distribution (OOD) evaluation (§4.2), which probes the robustness of the binary `mercado` construct across four external, heterogeneous PT-BR datasets.

In summary, our primary contributions are:

- **Standardised Multi-Family Protocol:** We formalize a rigorous comparative framework across three distinct paradigm families (TF–IDF linear models, PT-BR BERT encoders, and zero/few-shot LLMs) utilizing an identical 90/10 split, identical metrics, and Bonferroni-corrected pairwise McNemar significance tests.
- **Formulation-Independent Evaluation:** We analyze both binary and multiclass task structures on shared underlying documents, isolating the mathematical effect of the formulation from data-selection biases.
- **Leakage-Aware Ensembling:** We evaluate multiple combination strategies (majority voting and stacking meta-classifiers on an isolated internal partition) supported by token-level agreement analysis and a strict CV-only model selection funnel.
- **OOD Generalization and Cost Audit:** We provide an out-of-distribution framework transferring the selected systems onto four external PT-BR benchmarks [Monteiro et al. 2018, Garcia et al. 2022], mapping the direct relationship between covariate shift magnitude (MMD^2) and performance degradation, complemented by standardized compute cost and probability calibration cards.

2. Related Work

Automated news classification and triage within the financial and economic domains have seen significant growth, driven by the need to filter massive volumes of stream data before feeding downstream tasks like sentiment analysis, market impact prediction, or fact-checking [Araci 2019, Yang et al. 2020]. However, the adaptation of these upstream pipeline stages to Brazilian Portuguese (PT-BR) is still emerging and largely fragmented.

Within the PT-BR scientific ecosystem, research on misinformation and text categorization has primarily yielded generic, multi-domain datasets. Prominent examples in-

clude the *Fake.Br Corpus* [Monteiro et al. 2018], which focuses on political and celebrity news, and more recent alternatives like *FakeRecogna* [Garcia et al. 2022]. While these efforts provide essential infrastructure for general classification pipelines, they consistently treat economic news as a background theme rather than isolating it as a target for specific domain-filtering. Consequently, the nuanced vocabulary and implicit label shifts inherent to financial media have remained under-explored.

On the architectural spectrum, Portuguese NLP has evolved from sparse linear models to deep Transformer-based encoders and large language models (LLMs). The development of BERTimbau [Souza et al. 2020] established a robust baseline for general-purpose PT-BR understanding. This paved the way for specialized variants, such as FinBERT-PT-BR [Santos et al. 2023], which adapts the vocabulary to local financial contexts, and DeBERTa-style frameworks like DeB3RTa [Pires et al. 2025] that maximize structural representations. Concurrently, a native generation of PT-BR instruction-tuned LLMs has surfaced, featuring models like Sabiá [Pires et al. 2023], Bode [Garcia et al. 2024a], and Tucano [Corrêa et al. 2025].

Despite this diverse array of tools, the literature still lacks a systematic, rigorous comparison under controlled experimental conditions. Prior works often report performance across varying splits or disjoint evaluation metrics, neglecting the critical role of compute-to-performance trade-offs in operational triage environments. Our study fills this gap by executing a multi-family benchmark on unified, shared splits with strict significance and calibration tracking. Furthermore, we expand the conventional scope of the FolhaUOL collection by processing it as a balanced, multi-class editorial categorizer, a formulation that uncovers systemic label and covariate shifts when models are transitioned out-of-distribution.

3. Methodology

3.1. Corpus and Task

The corpus comprises Folha de São Paulo articles [Santana 2018] labelled by their editorial section. The *binary* task contrasts the economic section `mercado` against all others (`outros`); the *multiclass* task uses the seven highest-volume sections — `poder`, `colunas`, `mercado`, `esporte`, `mundo`, `cotidiano`, `ilustrada` — plus `outros`. Crucially, both formulations are built on the *same* document indices: only the label vector y changes, so every model family and both tasks see exactly the same documents in exactly the same roles, and differences between formulations are not confounded with sampling.

We use a single stratified split with a global seed. A frozen 10% test set is held out for final evaluation only — never used for tuning, threshold selection, early stopping, or model selection — and the remaining 90% forms the modelling pool, internally an $80/10$ train/validation split ($80/10/10$ overall). The validation tenth handles probability calibration, threshold and early-stopping decisions, and training of the stacking meta-classifier, and is never reported on its own; variance-aware in-domain estimates come instead from a stratified 5-fold cross-validation over the entire 90% pool. Since only the CV pool and the frozen test set are ever reported, we describe the protocol externally as 90/10. Stratification is applied throughout, and both the validation and the test partitions preserve the natural $\approx 12.5\%$ `mercado` prevalence.

3.2. Models, Tuning and Selection

We compare three model families under a single protocol.[1] The *TF–IDF* pipelines pair a `TfidfVectorizer` with a linear head (logistic regression, linear SVM, or multinomial naive Bayes), fitted jointly in an `sklearn` pipeline. The *BERT* encoders are the PT-BR BERTimbau [Souza et al. 2020], FinBERT-PT-BR [Santos et al. 2023], NorBERTo [Silva et al. 2026], and the DeBERTa-style DeB3RTa [Pires et al. 2025]; each is fine-tuned by a small per-model, per-task random search (5 trials over learning rate, batch size/accumulation, epochs, weight decay and warmup, with every trial scored on an inner validation split) at a fixed `max_length`$= 500$ and with early stopping on validation. The *LLMs* — Mistral-7B [Jiang et al. 2023] and Qwen2.5-7B [Yang et al. 2024] — instruction-tuned and prompted zero/few-shot with greedy decoding, receive the unmodified article text.

All reported systems use the same stopword-removed canonical generation: a versioned NLTK Portuguese list (207 terms) is removed before feature extraction, via the vectoriser for TF–IDF and upstream of subword tokenisation for BERT. For BERT this is a declared experimental condition rather than an innocuous normalisation, since the encoders were pretrained *with* stopwords; we adopt it for cross-family consistency. LLM prompts are the declared exception and receive natural text.

The three families are tuned at deliberately different resolutions. TF–IDF is searched with `RandomizedSearchCV` (60 trials over an inner stratified 5-fold kept independent of the reporting folds); BERT is tuned by a smaller custom random search (5 trials per model and task, scored on a single inner-validation split); and the LLMs are not searched at all. This asymmetry is a declared consequence of compute cost rather than an oversight, and we read the comparison accordingly, as best-of-family under a fixed budget. The primary metric is F_1 (binary, with the positive class `mercado`) and macro-F1 (multiclass); we also report precision, recall, accuracy, and PR-AUC (the area under the precision–recall curve, robust to the class imbalance), and, for calibrated probabilistic outputs, the Brier score and expected calibration error. PR-AUC, Brier and ECE are `N/A` for the LLMs, whose label-extraction yields deterministic scores. Pairwise differences are tested with McNemar under Bonferroni correction over the $K(K-1)/2$ family of comparisons.

Both formulations combine the same five base members: the three TF–IDF linear heads (logistic regression, linear SVM and multinomial naive Bayes) and the two strongest BERT encoders, BERTimbau and NorBERTo (Tables 1 and 2); FinBERT-PT-BR and DeB3RTa are not used as members. Over this pool we build a majority vote and a stacking meta-classifier, the latter fit on the internal calibration partition and never on the test set; to keep stacking comparable under the CV regime, we estimate it by stacked-CV over out-of-fold member probabilities. Model selection is then handled by a CV-only *selection funnel* that ranks candidates by the cross-validation mean of the primary metric, so the held-out test set plays no role. The funnel breaks Bonferroni-corrected McNemar ties by deployment cost and advances the best individual model and the best eligible ensemble

[1] Code and the per-model result cards behind every table (standardised metrics, cost measurements, and hyperparameter-search payloads, one card per model×task×evaluation regime) are available at `https://anonymous.4open.science/r/economy-classifier-anon-07E1/`. In case of any issues, please contact `almeida.diego@unifesp.br`.

of each formulation to the OOD probe of §4.2.

4. Results and Discussions

4.1. In-domain Results

Tables 1 and 2 report the binary and multiclass tasks for the three model families and the ensembles. The ordering is consistent across both formulations: the stacking ensemble is the strongest in-domain filter, leading on every class, with `esporte` the easiest section ($F_1 \approx 0.99$) and `outros` together with the target `mercado` the hardest ($F_1 \approx 0.91$–0.92). Per-class F_1, reported in full in the long version, follows the same ordering.

Table 1. Binary task (`mercado` vs. `outros`): 5-fold CV F1 on the 90% pool (mean $\pm$ std) and held-out test F1, precision, recall, and PR-AUC (area under the precision–recall curve, suited to the $\approx 12.5\%$ positive prevalence). Best per column in bold. *N/A*: PR-AUC is undefined for parsed-text LLMs (deterministic scores); coverage < 1 marks the LLM parsed subset, whose rows are not bolded.

Model	F1 (CV)	F1	Precision	Recall	PR-AUC	Coverage
TF–IDF + LogReg	0.8557 ± 0.0057	0.8608	0.8344	0.8889	0.9358	–
TF–IDF + LinearSVC	0.8593 ± 0.0037	0.8605	0.8971	0.8269	0.9406	–
TF–IDF + MultinomialNB	0.7751 ± 0.0015	0.7818	0.7212	0.8536	0.8224	–
BERTimbau	0.9044 ± 0.0025	0.9072	0.9222	**0.8927**	0.9632	–
NorBERTo	0.8994 ± 0.0034	0.9041	0.9188	0.8898	0.9621	–
FinBERT-PT-BR	0.9018 ± 0.0044	0.9005	0.9170	0.8846	0.9600	–
DeB3RTa	0.8300 ± 0.0027	0.8199	0.8650	0.7792	0.8986	–
Qwen2.5-7B (0-shot)	–	0.6337	0.5034	0.8549	*N/A*	0.9907
Qwen2.5-7B (few-shot)	–	0.6316	0.4832	0.9114	*N/A*	0.9764
Mistral-7B (0-shot)	–	0.5786	0.4662	0.7622	*N/A*	0.9835
Mistral-7B (few-shot)	–	0.6398	0.5532	0.7586	*N/A*	0.9518
Majority voting	0.8904 ± 0.0039	0.8984	0.9051	0.8918	–	–
Stacking	$\mathbf{0.9053 \pm 0.0043}$	**0.9119**	**0.9367**	0.8884	**0.9649**	–

Because the OOD funnel (§4.2) deploys the multiclass model collapsed to the `mercado`-vs-rest decision, we compare that projection — the *Proj.* F_1 column of Table 2, i.e. labelling a document `mercado` iff its multiclass argmax is `mercado` — against each system's dedicated binary head (Table 1) on the shared test indices. Collapsing never hurts: every system's projected F_1 matches or exceeds its binary head, with gains of $+0.004$ to $+0.025$. Under paired McNemar with Bonferroni correction (family-wise $\alpha = 0.05$), the gain is significant for all three TF–IDF models and the stacking ensemble ($+0.012$), but not for the BERT encoders or majority voting, whose binary and projected heads are statistically equivalent.

Figure 1 projects the frozen-BERTimbau embeddings of the *entire* corpus ($166{,}288$ documents) to two dimensions with a 2-D autoencoder bottleneck, in two variants that differ only in their loss. A label-supervised conditional autoencoder, whose classifier head organises the latent by section, recovers near-perfect angular separation of

Table 2. Multiclass task (top-7 editorial sections + `outros`): 5-fold CV macro-F1 on the 90% pool (mean ± std), held-out test macro-F1, the projected `mercado` F_1 (per-class F_1 of the argmax = `mercado` projection), and accuracy. Best per column in bold; coverage < 1 (LLM parsed subset) rows are not bolded.

Model	Macro-F1 (CV)	Macro-F1	Proj. F_1	Acc.	Cov.
TF–IDF + LogReg	0.8880 ± 0.0032	0.8956	0.8810	0.8921	–
TF–IDF + LinearSVC	0.8901 ± 0.0024	0.8989	0.8858	0.8957	–
TF–IDF + MultinomialNB	0.8010 ± 0.0030	0.8043	0.7917	0.8008	–
BERTimbau	0.9278 ± 0.0021	0.9316	0.9116	0.9299	–
NorBERTo	0.9263 ± 0.0020	0.9334	0.9166	0.9316	–
FinBERT-PT-BR	0.9258 ± 0.0022	0.9298	0.9045	0.9273	–
DeB3RTa	0.8621 ± 0.0031	0.8703	0.8570	0.8656	–
Qwen2.5-7B (0-shot)	–	0.4823	0.6112	0.5049	0.9872
Qwen2.5-7B (few-shot)	–	0.5717	0.6875	0.5849	0.8835
Mistral-7B (0-shot)	–	0.3854	0.4471	0.3970	0.9518
Mistral-7B (few-shot)	–	0.4359	0.4766	0.4609	0.7149
Majority voting	0.9116 ± 0.0027	0.9197	0.9079	0.9174	–
Stacking	$\mathbf{0.9361 \pm 0.0010}$	**0.9407**	**0.9235**	**0.9388**	–

the eight editorial sections (15-NN accuracy 0.996, cosine silhouette $+0.78$); the unsupervised common autoencoder (λ=0, reconstruction only) on the same embeddings does not (15-NN 0.713, cosine silhouette -0.09). Holding architecture, data and encoder fixed and toggling only the class-supervision term isolates the effect: the editorial signal is present in the frozen embedding but entangled, surfacing only once the projection is told the labels.

Figure 2 plots test-set performance against total wall-clock cost (training + inference, log scale; marker area $\propto$ parameters). The trade-off is steep: the TF–IDF models reach $F_1 \approx 0.86$ on CPU with sub-million parameters and the highest throughput ($\sim$500 docs/s), whereas the BERT encoders add only $\sim$0.05 F_1 at roughly two orders of magnitude more parameters and GPU fine-tuning. The stacking ensemble attains the best F_1 but at the additive cost of its five members, and the zero-/few-shot LLMs are inference-only yet the slowest ($\sim$1 doc/s) and least accurate (F_1 0.58–0.64), off the Pareto front.

4.2. Out-of-distribution Evaluation

Having selected the funnel systems on in-domain cross-validation, we now transfer them — frozen weights, vocabularies, decision threshold, and, for ensembles, combination rule — onto four heterogeneous PT-BR corpora: Fake.Br [Monteiro et al. 2018], FakeRecogna [Garcia et al. 2022], PortugueseNewsDataset [Garcia et al. 2024b], and Recogna-Summ [Paiola et al. 2024]. None carries Folha's editorial section space, so each corpus's native annotation is projected onto $\{\texttt{mercado}, \texttt{outros}\}$ and the evaluation is binary. The binary classifiers transfer directly; for a multiclass head, the projection of Eq. (1) maps it onto the same target, for class posteriors $p(c \mid x)$,

$$\hat{y}(x) = \mathbf{1}\Big[\arg\max_{c} p(c \mid x) = \texttt{mercado}\Big], \qquad s(x) = p(\texttt{mercado} \mid x), \tag{1}$$

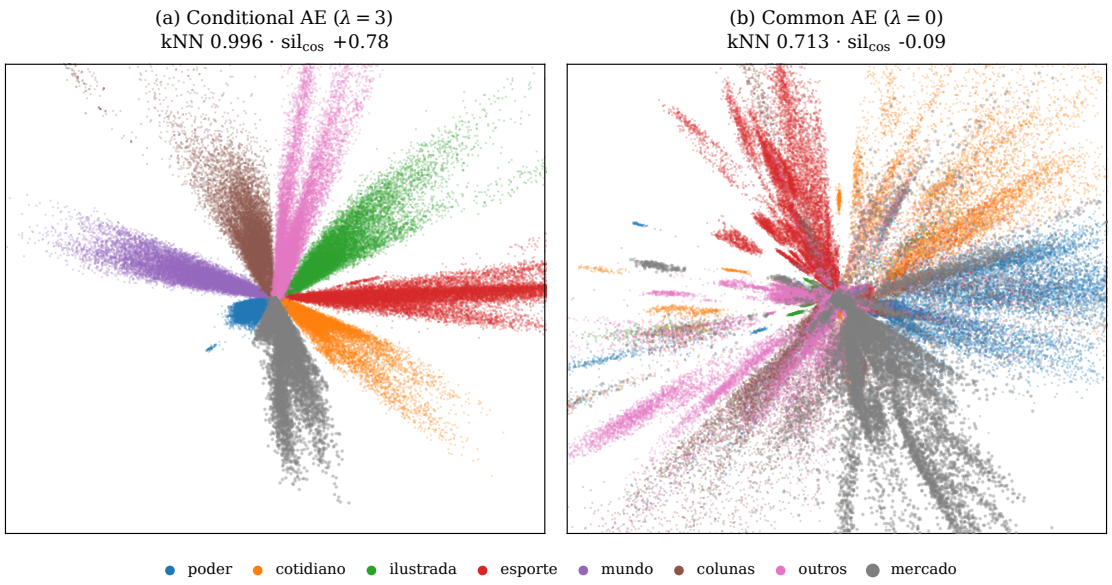


Figure 1. Two-dimensional autoencoder projection of the frozen BERTimbau embeddings of the *entire* Folha-SP corpus (166,288 documents, train+val+test), coloured by editorial section. (a) A label-supervised conditional autoencoder (λ=3) separates the eight sections into distinct angular rays; (b) the unsupervised common autoencoder (λ=0, reconstruction only) on the same embeddings mixes them. Panel annotations: 15-NN accuracy and cosine silhouette against the editorial label (the star-shaped latent makes Euclidean silhouette uninformative). The binary target `mercado` is shown in grey.

so both formulations face the same targets. No per-corpus threshold re-calibration is applied.

To interpret each degradation, we first quantify how far each corpus sits from the Folha-SP training pool in a frozen BERTimbau embedding space (`max_length`= 256), using a kernel two-sample test [Gretton et al. 2012] (MMD^2 with a permutation p-value). All four corpora reject the null ($p < 10^{-3}$), whereas a Folha self-control does not (MMD$^2 \approx 0.001$, $p = 0.25$), confirming that the measured distance reflects genuine cross-corpus shift rather than sampling noise.

One caveat is specific to RecognaSumm, the corpus closest in register to the training source. Our training corpus is Folha de São Paulo (the *FolhaUOL* collection, `folha.uol.com.br`), sampled from January 2015 to September 2017, whereas RecognaSumm is a multi-outlet corpus (G1, CNN Brasil, O Globo, UOL, among others) spanning 2013–2023. It does *not* contain Folha de São Paulo itself: its only overlap with the training publisher is the sibling UOL portal (`noticias.uol.com.br`/`economia.uol.com.br`) under the same Grupo Folha umbrella, and those articles concentrate in 2021–2023. Even this nearest corpus thus bundles a publisher-but-not-same-portal difference with a near-disjoint temporal window, so its degradation is not a same-source control — reinforcing the temporally-aware analysis flagged in §5.

Table 3 reports the F_1 on the `mercado` target per corpus for the canonical binary

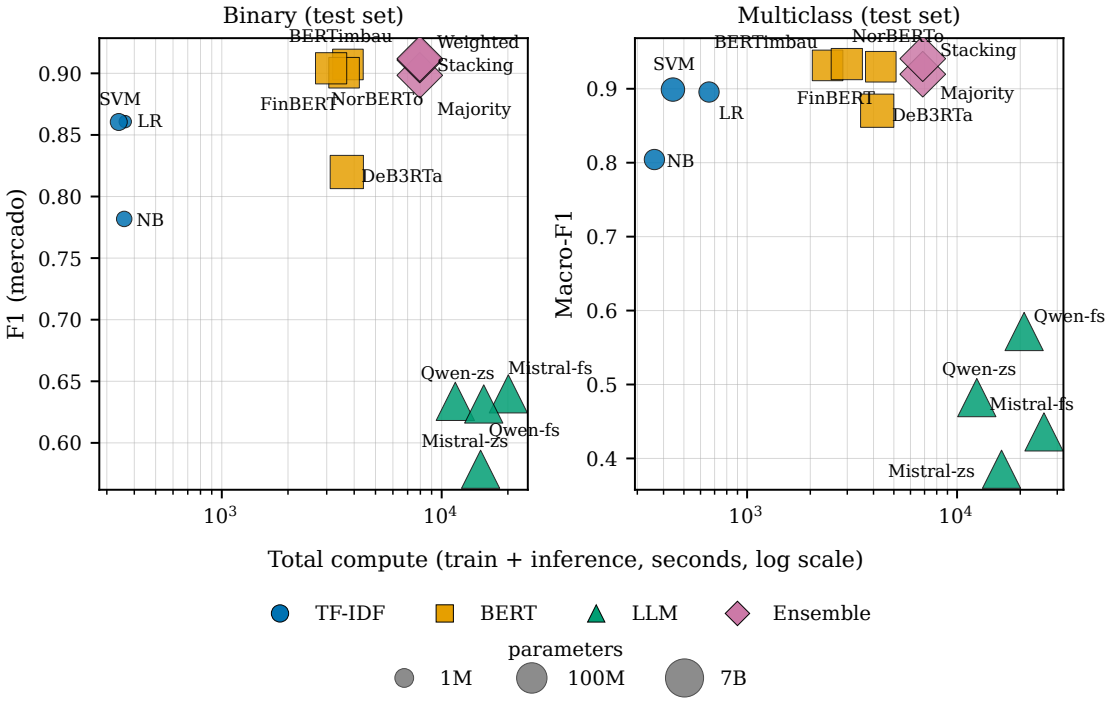


Figure 2. Test-set performance versus total cost (train + inference seconds, log scale): binary (left, F_1) and multiclass (right, macro-F1). Marker area $\propto$ parameter count; LLMs (triangles) are inference-only.

systems.

Table 3. Out-of-distribution transfer of the canonical (`nostop`) binary classifiers, applied directly to each corpus: F_1 on the `mercado` target, inference at `max_length`=500. F_1^{in} is the in-domain test anchor (Table 1); best per corpus in bold.

System	F_1^{in}	Fake.Br	FakeRecogna	PortNews	RecognaSumm
TF–IDF + LogReg	0.8608	0.1160	0.1847	0.5783	0.5807
TF–IDF + LinearSVC	0.8605	0.1353	0.1953	0.4985	0.5848
TF–IDF + MultinomialNB	0.7818	**0.1516**	0.3683	**0.6409**	**0.6264**
BERTimbau	0.9072	0.1310	**0.4208**	0.5473	0.5598
FinBERT-PT-BR	0.9005	0.1230	0.4102	0.6352	0.6004
NorBERTo	0.9041	0.1281	0.3248	0.5662	0.6241

Figure 3 relates the relative F_1 degradation of the canonical systems to **(a)** positive-class prevalence and **(b)** covariate shift (MMD^2) per corpus.

5. Conclusion

We presented a standardised, calibration- and significance-aware protocol for filtering economic news in Brazilian Portuguese, spanning TF–IDF, PT-BR BERT encoders, LLMs and ensembles across two shared-split formulations, with a CV-only selection funnel feeding a zero-shot OOD probe over four corpora. Ensembles give the strongest in-domain filter and BERT leads among individual models, but the cost–benefit and OOD

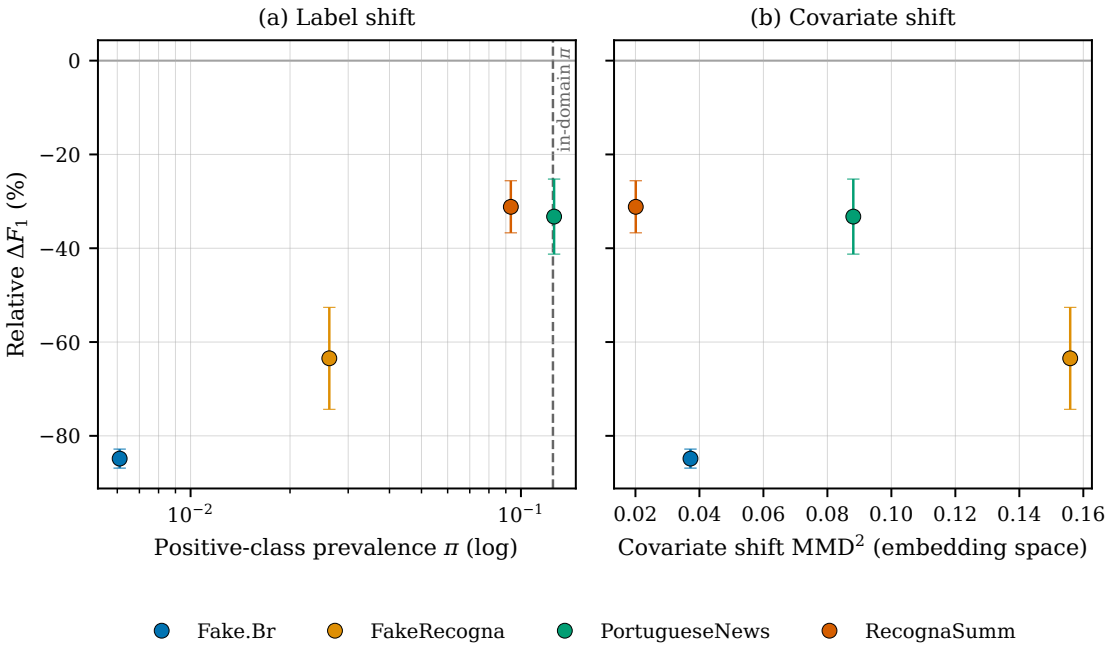


Figure 3. Relative ΔF_1 (mean over the canonical nostop systems, ± 1 std) against (a) positive-class prevalence π (dashed line $=$ in-domain $\pi \approx 0.125$) and (b) covariate shift MMD^2 in a frozen BERTimbau embedding space (`max_length`$= 256$).

evidence qualify that headline. A TF–IDF model recovers most of the performance at one-to-two orders of magnitude less cost and stays competitive on the editorially-closest out-of-domain corpora, which is the trade-off that matters for an upstream-triage stage.

Our findings are nevertheless constrained by several factors. The `mercado` label is an editorial category rather than a topical annotation: it records where the newsroom filed an article, so overlapping content filed elsewhere is irreducible label noise. The multiclass formulation makes this visible rather than removing it, and the heterogeneous `colunas` section, defined by *format* rather than topic, resists topical classification by construction. The corpus is also single-source (Folha de São Paulo), so the in-domain results do not measure cross-newsroom generalisation; the OOD probe, in turn, differs from the source in genre, register, length and prevalence simultaneously, which is why each delta is paired with an explicit shift measurement. Three further constraints are declared. LLM scores are deterministic, so PR-AUC, Brier and ECE are `N/A` and low-coverage rows form a separate McNemar family. The tuning budget is asymmetric — TF–IDF searched at 60 trials, BERT by a coarser 5-trial search, and LLMs unsearched — so the comparison is best-of-family under a fixed budget. Finally, the stacking CV estimate is the mildly optimistic stacked-CV, with full nested CV left for future work.

Beyond closing that gap, future work includes cross-newsroom transfer with editorial labels, PT-BR-specialised LLMs, transfer of the complete multiclass label space, cross-corpus normalisation strategies that disentangle genuine distribution shift from the length, genre and prevalence confounds it currently bundles together, and temporally-aware analyses that model concept drift across the news stream rather than treating the corpus as time-invariant.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this work, the authors used AI-assisted tools and generative artificial intelligence technologies: Gemini to polish the phrasing, enhance grammatical readability, and translate draft text into structured English prose, and Claude Code (Anthropic), an agentic LLM-based coding assistant, to help generate and refactor parts of the experimental code. After using these services, the authors thoroughly reviewed, edited, and verified all generated content, and take full responsibility for the final scientific accuracy.

Acknowledgments

The authors would like to thank Serasa Experian, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Research Grants 2021/14725-3, 2024/17511-2, and 2025/21207-0, for supporting the development of this study.

References

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint*.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2025). Tucano: Advancing neural text generation for Portuguese. *Patterns*, 6(6):101325.

Garcia, G. L., Afonso, L. C. S., and Papa, J. P. (2022). FakeRecogna: A new Brazilian corpus for fake news detection. In *Computational Processing of the Portuguese Language (PROPOR 2022)*, volume 13208 of *Lecture Notes in Computer Science*, pages 57–67. Springer.

Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Candido Júnior, A., Jodas, D. S., Afonso, L. C. S., Guilherme, I. R., Penteado, B. E., and Papa, J. a. P. (2024a). Introducing bode: A fine-tuned large language model for Portuguese prompt-based task. *arXiv preprint*.

Garcia, K., Shiguihara, P., and Berton, L. (2024b). Breaking news: Unveiling a new dataset for Portuguese news classification and comparative analysis of approaches. *PLOS ONE*, 19(1):e0296929.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7B. *arXiv preprint*.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in Portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language (PROPOR)*, pages 324–334. Springer.

Paiola, P. H., Garcia, G. L., Jodas, D. S., Correia, J. V. M., Sugi, L. A., and Papa, J. P. (2024). RecognaSumm: A novel Brazilian summarization dataset. In *Proceedings of*

the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024), pages 575–579, Santiago de Compostela, Spain. Association for Computational Linguistics.

Pires, H., Paucar, L., and Carvalho, J. a. P. (2025). DeB3RTa: A transformer-based model for the Portuguese financial domain. *Big Data and Cognitive Computing*, 9(3):51.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Intelligent Systems — 12th Brazilian Conference (BRACIS)*, volume 14197 of *Lecture Notes in Computer Science*. Springer.

Santana, M. (2018). News of the site Folha de São Paulo (Brazilian Newspaper). `https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol`. Accessed 2026-05-21.

Santos, L. L., Bianchi, R. A. C., and Costa, A. H. R. (2023). FinBERT-PT-BR: Análise de sentimentos de textos em português do mercado financeiro. In *Anais do II Brazilian Workshop on Artificial Intelligence in Finance (BWAIF)*, pages 144–155. SBC.

Silva, E. S. N., Costa, P. B., Vlasman, R. C., Costa, R. P., Silva, H. L. P., Pellicer, L. F. A. O., Rinaldo, G., Almeida, R. A., Rabbani, D. S. R., Oestreich, C. O., and Caridá, V. F. (2026). NorBERTo: A ModernBERT model trained for Portuguese with 331 billion tokens corpus. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 183–193, Salvador, Brazil. Association for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems — 9th Brazilian Conference (BRACIS)*, volume 12320 of *Lecture Notes in Computer Science*. Springer.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. (2024). Qwen2.5 Technical Report. *arXiv preprint*.

Yang, Y., Uy, M. C. S., and Huang, A. (2020). FinBERT: A pretrained language model for financial communications. *arXiv preprint*.