A Multi-Corpus Benchmark of Classical Fake News Classifiers with Contextual Portuguese Embeddings

Adriel L. V. Mori[1,2], Eliomar Araújo Lima[1,2], Valdemar V. Graciano Neto[1,2], Jacson R. Barbosa[1,2], Rogerio Rodrigues Carvalho[1,2], Arlindo R. Galvão Filho[1,2]

[1]Institute of Informatics – Federal University of Goiás (UFG)

[2]Advanced Knowledge Center for Immersive Technologies (AKCIT)
Goiânia – GO – Brazil

adrielmori@discente.ufg.br,

{eliomar,valdemarneto,arlindogalvao,jacson_rodrigues,rogerior}@ufg.br

Abstract. *This paper presents a systematic benchmark for fake news detection in Brazilian Portuguese, combining multiple datasets, four contextual embedding models, and eight classical supervised classifiers under a unified evaluation protocol. The results show that, when a strong classifier such as* `SVC-RBF` *is adopted, performance varies more across embedding models than across classifiers. Among the evaluated encoders,* `albertina_ptbr_900m` *achieved the best overall generalization on the test set, while* `bertimbau_large` *showed the strongest validation performance. Overall, the study highlights representation quality as a central factor for robust classical fake news classification in Brazilian Portuguese.*

1. Introduction

Fake news detection has become an important research problem in digital communication, especially in contexts where misleading content affects public debate, health communication, and political discourse [Vosoughi et al. 2018]. In Brazilian Portuguese, this challenge has motivated the creation of several annotated corpora, such as `Fake.Br`, `COVID19.BR`, and `FactCK.Br`, which have supported the development of automatic detection methods under distinct domains, annotation schemes, and linguistic profiles [Monteiro et al. 2018, Martins et al. 2021, Moreno and Bressan 2019]. However, despite the growing number of available datasets, the empirical landscape remains fragmented: many studies focus on a single corpus, a limited experimental setup, or a restricted family of models, which makes broader comparative conclusions difficult.

At the same time, recent advances in Portuguese-specific language modeling have substantially improved the quality of contextual text representations. Models such as BERTimbau and Albertina provide pretrained encoders specifically adapted to Portuguese, enabling richer semantic embeddings for downstream tasks [Souza et al. 2020, Santos et al. 2024]. This creates a particularly relevant scenario for classical supervised learning. While modern fake news research often emphasizes end-to-end neural architectures or large language models, classical classifiers remain attractive because they are computationally efficient, easier to interpret, and still highly competitive when combined with strong document representations. In this setting, an important open question is whether

performance gains are driven primarily by the classifier itself or by the semantic quality of the embedding space.

This paper addresses that question through a large-scale and systematic benchmark of classical fake news classifiers in Brazilian Portuguese. The study combines multiple corpora with different class balances, document lengths, and topical domains, four contextual embedding models specifically designed for Portuguese, and eight classical supervised classifiers under a unified evaluation protocol. This design allows a broader analytical comparison than is usually reported in the literature, since it evaluates the behavior of the same classifier–embedding combinations across heterogeneous datasets rather than under a single isolated corpus. From this perspective, the innovation of the study lies not in proposing a new architecture, but in offering a controlled multi-corpus analysis of how representation quality influences classical classification performance in Portuguese fake news detection.

The central hypothesis is that, once a sufficiently strong classifier is adopted, especially among margin-based nonlinear models, the main source of performance variation becomes the embedding model rather than the downstream classifier. Thus, the contribution of this work is both empirical and methodological: empirically, by providing a broad comparison between Portuguese embedding families across several fake news datasets; and methodologically, by showing that advances in semantic representation can be a decisive factor for robust classical pipelines. As shown in the results, this effect is consistent not only in the best-performing systems, but also in the average behavior across classifiers, reinforcing the practical relevance of contextual Portuguese embeddings for multi-dataset fake news classification.

2. Related Work

Research on fake news detection has expanded substantially in recent years, motivated by the social and political impact of online misinformation and by the growing availability of annotated corpora. A recent systematic literature review by Villela et al. [Villela et al. 2023] shows that the area has evolved from traditional machine learning pipelines based on lexical and stylistic features to more complex deep learning and hybrid approaches. At the same time, the review highlights recurring limitations in the literature, including the predominance of controlled datasets, the concentration on English, and the lack of broader comparative protocols across corpora and representation methods. These limitations are particularly relevant in Brazilian Portuguese, where the number of available datasets has increased, but empirical comparisons are still fragmented.

In Brazilian Portuguese, one of the earliest and most influential resources is *Fake.Br*, introduced by Monteiro et al. [Monteiro et al. 2018], which established a benchmark corpus and baseline results for fake news detection. Building on this dataset, Silva et al. [Silva et al. 2020] conducted a comprehensive study of textual representations and classical classifiers, showing that bag-of-words and linguistic features can outperform Word2Vec- and FastText-based representations in that setting. Jeronimo et al. [Jeronimo et al. 2019] proposed a complementary perspective based on subjective language, modeling fake news detection through semantic distances to subjectivity lexicons and emphasizing robustness under cross-domain and cross-source evaluation. Paixão et al. [Paixão et al. 2020] further extended the analysis on Fake.Br by comparing classical

and deep learning methods under different feature sets and by complementing classification with topic modeling, reinforcing the value of corpus-level qualitative analysis.

Other studies in Portuguese explored adjacent modeling perspectives. De Morais et al. [de Morais et al. 2020] moved beyond the binary setting by distinguishing among fake, satirical, objective, and legitimate news, showing that fake news detection can benefit from richer label structures. Gôlo et al. [Gôlo et al. 2023] investigated one-class learning with multimodal variational autoencoders, combining textual and topic-based information to improve representation learning in low-label scenarios. Sousa et al. [Sousa et al. 2022] combined convolutional neural networks and machine learning algorithms for Portuguese fake news detection, while Pires and Guerreiro e Silva [Pires and Guerreiro e Silva 2024] evaluated BERT, mBERT, and BERTimbau, reporting gains for Portuguese-specific transformers. Taken together, these studies indicate that the literature in Portuguese has progressively moved from sparse lexical features to richer semantic and contextual representations.

Despite these advances, most prior studies still evaluate a limited number of datasets, focus on a single representation family, or compare only a small set of classifiers. This makes it difficult to disentangle whether performance gains come from the classifier, the representation, or the particular properties of the dataset being used. In addition, much of the literature in Portuguese remains centered on single-corpus analyses, especially on Fake.Br, which is methodologically useful but insufficient to characterize the robustness of a method across different textual domains, annotation practices, and levels of class imbalance.

Our work differs from this line of research in a specific and complementary way. Rather than proposing a new neural architecture, a new handcrafted feature space, or a single-dataset improvement, we provide a controlled and systematic benchmark of classical supervised classifiers over contextual Portuguese embeddings across multiple fake news datasets. This design allows us to analyze, in a unified setting, how much performance depends on the downstream classifier and how much is driven by the embedding space itself. From this perspective, the main differential of our approach is analytical: we shift the focus from isolated model gains to a broader comparison of representation quality, classifier behavior, and cross-dataset robustness in Brazilian Portuguese fake news detection.

3. Materials and Methods

3.1. Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a binary text classification dataset, where x_i is a news item written in Brazilian Portuguese and $y_i \in \{0, 1\}$ is the veracity label. In this benchmark, $y_i = 1$ indicates *REAL* and $y_i = 0$ indicates *FAKE*. The objective is to evaluate the predictive performance of classical classifiers when trained on dense document embeddings extracted from different Portuguese language encoders.

For each encoder f_θ, a document x_i is mapped into a fixed-dimensional representation $\mathbf{z}_i \in \mathbb{R}^{d_\theta}$. In this work, $\mathbf{z}_i$ is obtained from the encoder token-level representations using attention-mask-aware mean pooling, so that only valid non-padding tokens contribute to the final document vector. The complete embedding extraction procedure, including the encoder choices, tokenization setup, maximum sequence length, and pooling definition, is detailed in Section 3.3.

Given the resulting fixed embedding $\mathbf{z}_i$, each classical classifier learns a mapping

$$g_\phi : \mathbb{R}^{d_\theta} \rightarrow \{0, 1\}, \tag{1}$$

so that the final prediction is given by

$$\hat{y}_i = g_\phi(\mathbf{z}_i). \tag{2}$$

After the embeddings are generated, no additional pooling is performed during classifier training; the classical models operate only on the precomputed document vectors.

3.2. Datasets

The benchmark is composed of multiple Brazilian Portuguese and Portuguese-language fake news datasets covering different domains, annotation protocols, and textual granularities, including BOATOSBR [Cantarino 2024], CENTRALFATOS [Marques et al. 2022], COVID19BR [Martins et al. 2021], FACTCKBR [Moreno and Bressan 2019], FAKEBR [Monteiro et al. 2018], FAKETRUEBR [Chavarro et al. 2023], FCN [Santos 2022], FNEWSSET [da Silva et al. 2020], FRECOGNA [Garcia et al. 2022], MUMINPT [Nielsen and McConville 2022], and TRE300 [Gôlo et al. 2024]. This heterogeneity is methodologically relevant because it exposes the classifiers to substantial variation in topic, writing style, class balance, and document length.

Table 1 summarizes the datasets used in the benchmark. The collection includes both approximately balanced and highly imbalanced corpora. For example, FAKEBR, FAKETRUEBR, FCN, FNEWSSET, FRECOGNA, and TRE300 are approximately balanced, whereas CENTRALFATOS, FACTCKBR, and MUMINPT are strongly skewed toward the *FAKE* class. Text length also varies substantially across corpora: some datasets contain very short textual items, such as FNEWSSET and MUMINPT, while others contain longer documents, such as CENTRALFATOS, FCN, and the *REAL* portion of TRE300. This variability makes the benchmark suitable for evaluating the robustness of embedding-based classical models under heterogeneous linguistic conditions.

Table 1. Descriptive statistics of the datasets used in the benchmark. Avg. tokens are reported separately for the *FAKE* and *REAL* classes.

Dataset	FAKE	REAL	Total	Avg. tokens (FAKE)	Avg. tokens (REAL)
BOATOSBR [Cantarino 2024]	1,888	1,516	3,404	135.13	611.94
CENTRALFATOS [Marques et al. 2022]	10,282	34	10,316	649.50	373.21
COVID19BR [Martins et al. 2021]	905	1,494	2,399	172.16	104.85
FACTCKBR [Moreno and Bressan 2019]	1,866	240	2,106	42.24	35.68
FAKEBR [Monteiro et al. 2018]	3,600	3,600	7,200	182.75	184.30
FAKETRUEBR [Chavarro et al. 2023]	1,791	1,791	3,582	156.12	507.11
FCN [Santos 2022]	1,044	1,020	2,064	521.88	490.12
FNEWSSET [da Silva et al. 2020]	300	300	600	18.94	24.20
FRECOGNA [Garcia et al. 2022]	5,951	5,951	11,902	101.90	68.55
MUMINPT [Nielsen and McConville 2022]	1,339	65	1,404	18.75	16.35
TRE300 [Gôlo et al. 2024]	148	152	300	27.63	659.30

Before training, all labels were normalized to the binary scheme {*FAKE*, *REAL*}. No manual rebalancing was imposed, so that each corpus preserved its original class distribution. This design choice allows the evaluation to reflect the intrinsic difficulty of each dataset rather than an artificially homogenized setting.

3.3. Embedding Models

To instantiate the representation step described in Section 3.1, four Portuguese encoders were used to generate dense document representations: `bertimbau_base`[1] and `bertimbau_large`[2], both from the BERTimbau family [Souza et al. 2020]; `albertina_ptbr_100m`[3], based on the Albertina PT-* family [Santos et al. 2024]; and `albertina_ptbr_900m`[4], also from the Albertina PT-* family [Rodrigues et al. 2023]. These encoders were selected because they are specifically adapted to Portuguese and provide contextualized token representations suitable for downstream classification.

Given a document x_i, the tokenized input was truncated or padded to a maximum length of 128 tokens and processed by the selected encoder f_θ. The encoder produces a contextual representation matrix

$$\mathbf{H}_i = f_\theta(x_i) = \begin{bmatrix} \mathbf{h}_{i1}^\top \\ \mathbf{h}_{i2}^\top \\ \vdots \\ \mathbf{h}_{iT_i}^\top \end{bmatrix} \in \mathbb{R}^{T_i \times d_\theta}, \tag{3}$$

where $T_i \leq 128$ is the tokenized length and d_θ is the encoder-specific hidden dimensionality.

The document-level embedding is obtained using attention-mask-aware mean pooling over the last hidden states of the encoder. Let m_{it} be the attention-mask value for token t, where $m_{it} = 1$ indicates a valid token and $m_{it} = 0$ indicates a padding token. The document embedding is computed as

$$\mathbf{z}_i = \frac{\sum_{t=1}^{T_i} m_{it}\mathbf{h}_{it}}{\sum_{t=1}^{T_i} m_{it}} \in \mathbb{R}^{d_\theta}. \tag{4}$$

Thus, the method does not use the `[CLS]` representation or the encoder's `pooler_output`; instead, it computes a manual mean pooling over valid token embeddings.

In all experiments, embeddings were extracted in batches of size 16 to ensure a consistent inference setup across encoder families. The resulting embedding matrix for a dataset with N documents is

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^\top \\ \mathbf{z}_2^\top \\ \vdots \\ \mathbf{z}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times d_\theta}. \tag{5}$$

[1]Model artifact available at: `https://huggingface.co/neuralmind/bert-base-portuguese-cased`. Accessed on May 15, 2026.

[2]Model artifact available at: `https://huggingface.co/neuralmind/bert-large-portuguese-cased`. Accessed on May 15, 2026.

[3]Model artifact available at: `https://huggingface.co/PORTULAN/albertina-100m-portuguese-ptbr-encoder`. Accessed on May 15, 2026.

[4]Model artifact available at: `https://huggingface.co/PORTULAN/albertina-900m-portuguese-ptbr-encoder-brwac`. Accessed on May 15, 2026.

3.4. Classical Classification Models

Eight classical classifiers were evaluated on top of the document embeddings: Logistic Regression, Linear Support Vector Machine, RBF-kernel Support Vector Machine, Random Forest, Gradient Boosting, Gaussian Naive Bayes, k-Nearest Neighbors, and Decision Tree [Hosmer et al. 2013, Cortes and Vapnik 1995, Breiman 2001, Friedman 2001, Zhang 2004, Cover and Hart 1967, Breiman et al. 1984].

Let $\mathbf{z}_i \in \mathbb{R}^d$ denote the embedding of document i. Logistic Regression models the posterior probability of the positive class as

$$P(y_i = 1 \mid \mathbf{z}_i) = \sigma(\mathbf{w}^\top \mathbf{z}_i + b), \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function.

Linear SVM estimates a maximum-margin hyperplane by solving

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \tag{7}$$

subject to

$$y_i(\mathbf{w}^\top \mathbf{z}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0. \tag{8}$$

For the nonlinear SVM, the decision function is defined through the radial basis kernel

$$K(\mathbf{z}_i, \mathbf{z}_j) = \exp\left(-\gamma\|\mathbf{z}_i - \mathbf{z}_j\|^2\right). \tag{9}$$

Random Forest predicts by majority vote over M trees,

$$\hat{y}_i = \operatorname{mode}\left\{T_m(\mathbf{z}_i)\right\}_{m=1}^{M}, \tag{10}$$

whereas Gradient Boosting builds an additive ensemble

$$F_M(\mathbf{z}) = \sum_{m=1}^{M} \nu\, h_m(\mathbf{z}), \tag{11}$$

with learning rate ν and weak learners $h_m(\cdot)$.

Gaussian Naive Bayes assumes conditional independence among embedding dimensions given the class:

$$P(y = c \mid \mathbf{z}) \propto P(y = c) \prod_{j=1}^{d} \mathcal{N}(z_j \mid \mu_{cj}, \sigma_{cj}^2). \tag{12}$$

KNN predicts by majority vote among the k nearest neighbors:

$$\hat{y}_i = \operatorname{mode}\left\{y_j : \mathbf{z}_j \in \mathcal{N}_k(\mathbf{z}_i)\right\}. \tag{13}$$

Finally, the Decision Tree recursively partitions the feature space by minimizing an impurity criterion such as the Gini index

$$G(t) = 1 - \sum_{c\in\{0,1\}} p(c \mid t)^2. \tag{14}$$

3.5. Model Configuration and Hyperparameter Search

Hyperparameters were selected by exhaustive grid search using predefined search spaces for each classical classifier. The same configuration protocol was applied to all encoder–classifier pairs defined in the experimental design described in Section 3.6.

For classifiers sensitive to feature scale, namely Logistic Regression, LinearSVC, RBF-SVC, GaussianNB, and KNN, embeddings were standardized using `StandardScaler`. Tree-based models, namely Random Forest, Gradient Boosting, and Decision Tree, were trained directly on the raw embeddings.

The search spaces were defined as follows:

- **Logistic Regression:** $C \in \{0.01, 0.1, 1.0, 10.0\}$, solver $\in \{\texttt{liblinear}, \texttt{lbfgs}\}$
- **LinearSVC:** $C \in \{0.01, 0.1, 1.0, 10.0\}$
- **RBF-SVC:** $C \in \{1.0, 10.0\}$, $\gamma = \texttt{scale}$
- **Random Forest:** number of trees $\in \{200, 500\}$, max depth $\in \{\texttt{None}, 20\}$
- **Gradient Boosting:** number of estimators $\in \{100, 200\}$, learning rate $= 0.1$, max depth $= 3$
- **GaussianNB:** `var_smoothing` $\in \{10^{-9}, 10^{-8}\}$
- **KNN:** $k \in \{3, 5, 7, 11\}$, weights $\in \{\texttt{uniform}, \texttt{distance}\}$
- **Decision Tree:** max depth $\in \{\texttt{None}, 10, 20, 30\}$, min samples split $\in \{2, 5, 10\}$, criterion $\in \{\texttt{gini}, \texttt{entropy}\}$

3.6. Evaluation Design

The experimental design combines four embedding models with eight classical classifiers across multiple datasets, yielding a factorial evaluation over encoder–classifier pairs under heterogeneous corpus conditions. All stochastic procedures were executed with a fixed random seed of 42.

The data were partitioned into training, validation, and test subsets using the proportions

$$|\mathcal{D}_{\text{test}}| = 0.15|\mathcal{D}|, \qquad |\mathcal{D}_{\text{val}}| = 0.15|\mathcal{D}|, \qquad |\mathcal{D}_{\text{train}}| = 0.70|\mathcal{D}|, \tag{15}$$

with $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}} = \mathcal{D}$ and pairwise disjoint subsets.

To preserve the original distribution of both corpora and classes, the partitions were generated through stratification over the composite variable

$$s_i = (\texttt{dataset\_id}_i,\ y_i), \tag{16}$$

so that each split approximately preserves the joint distribution of dataset membership and label. This procedure is particularly important in a multi-corpus benchmark, because it avoids distortions caused by dataset-specific imbalance and ensures that train, validation, and test sets remain comparable.

Hyperparameter selection followed the search spaces described in Section 3.5. Model selection was performed using the validation set, and final performance was reported only on the held-out test set. Because the benchmark includes strongly imbalanced corpora, evaluation prioritized macro-averaged measures, especially macro-F_1, complemented by class-wise precision and recall.

4. Reproducibility

All code, configurations, and instructions for obtaining and preprocessing the corpora, subject to their original licenses, are available in the GitHub repository[5].

5. Results

We evaluated 32 classical pipelines, resulting from the combination of four embedding models and eight classifiers. The analysis below focuses exclusively on classical models and excludes zero-shot LLM-based systems. Since the main goal of this section is to assess the contribution of the embedding space, we report four complementary perspectives: (i) the best configuration obtained for each embedding, (ii) the aggregate average performance of each embedding across classifiers, (iii) the classifier-level breakdown, and (iv) the dataset-level robustness analysis. The first two perspectives preserve the aggregate benchmark view, while the latter two make explicit how performance varies across classifiers and heterogeneous corpora.

Table 2. Best downstream configuration for each embedding model, reporting validation and test performance of the best associated classifier.

Embedding	Best Clf.	CV Macro-F_1	Val Acc.	Val Macro-F_1	Test Acc.	Test Macro-F_1
albertina_ptbr_900m	svc_rbf	0.9196	0.9363	0.9293	**0.9376**	**0.9312**
bertimbau_large	svc_rbf	**0.9211**	**0.9435**	**0.9373**	0.9368	0.9305
bertimbau_base	svc_rbf	0.9186	0.9371	0.9302	0.9352	0.9285
albertina_ptbr_100m	svc_rbf	0.9020	0.9261	0.9181	0.9252	0.9174

Table 2 presents the best-performing classifier obtained for each embedding model under the aggregate test evaluation. In all four cases, the best configuration was achieved with the RBF-kernel SVM, suggesting that this classifier was particularly effective when applied to dense contextual embeddings. Under this peak-performance perspective, albertina_ptbr_900m obtained the highest aggregate test performance, reaching 0.9376 accuracy and 0.9312 macro-F_1. Although bertimbau_large achieved the highest validation scores, its aggregate test performance was only marginally below that of albertina_ptbr_900m. This narrow gap indicates that the strongest encoders produced highly competitive representations when paired with the same nonlinear classifier.

Table 3. Aggregate average test macro-F_1 of each embedding across all eight classical classifiers..

Embedding	Mean Test Macro-F_1
albertina_ptbr_900m	**0.8278**
bertimbau_large	0.8190
bertimbau_base	0.8145
albertina_ptbr_100m	0.8020

Table 3 provides a complementary aggregate view by averaging the test macro-F_1 of each embedding across all eight classifiers. From this perspective, albertina_ptbr_900m also achieved the highest mean test macro-F_1, followed by bertimbau_large, bertimbau_base, and albertina_ptbr_100m. This suggests that the

[5]https://github.com/adrielmori/wics-csbd2026

strongest embedding was not only effective in its best configuration, but also competitive across different classifier families. However, because the differences among the top embeddings are small, this result should be interpreted as descriptive evidence rather than definitive statistical superiority.

5.1. Classifier-Level Breakdown

Table 4. Mean macro-F_1 across the ten datasets (test split), per embedding–classifier pair; last column: average across the four embeddings.

Classifier	albertina_100m	albertina_900m	bertimbau_base	bertimbau_large	Mean
svc_rbf	0.7914	0.8219	0.8065	**0.8243**	**0.8110**
knn	0.7155	0.7389	0.7346	0.7439	0.7332
linearsvc	0.6843	0.7479	0.6954	0.7339	0.7154
logreg	0.6890	0.7381	0.6941	0.7313	0.7131
random_forest	0.7070	0.7114	0.7110	0.7158	0.7113
gradient_boosting	0.6952	0.7067	0.6684	0.6915	0.6905
decision_tree	0.6173	0.6315	0.6066	0.5902	0.6114
gaussian_nb	0.3922	0.3867	0.4025	0.3939	0.3938

Table 4 expands the aggregate analysis by reporting the individual behavior of the eight classical classifiers. The RBF-kernel SVM achieved the highest average dataset-level macro-F_1 across embeddings (0.8110), followed by KNN (0.7332), LinearSVC (0.7154), Logistic Regression (0.7131), and Random Forest (0.7113). This confirms that svc_rbf was not only selected as the best classifier in the aggregate embedding comparison, but also remained the strongest classifier when performance was averaged across datasets and embedding families.

The per-classifier breakdown also qualifies the interpretation of the embedding effect. Larger or stronger encoders tended to improve several classifiers, especially LinearSVC and Logistic Regression, but the magnitude of the gains was not uniform across all learning algorithms. For instance, albertina_ptbr_900m improved over albertina_ptbr_100m by $+6.36$ p.p. for LinearSVC and $+4.91$ p.p. for Logistic Regression, whereas the corresponding gain for Random Forest was only $+0.44$ p.p. This indicates that the benefit of richer contextual embeddings depends on the downstream classifier, and that the benchmark should be interpreted as an encoder–classifier interaction rather than as an embedding-only effect.

The classifier results also show that the weaker models were consistently less suitable for this benchmark. In particular, GaussianNB achieved the lowest average macro-F_1 (0.3938), suggesting that the distributional assumptions of this classifier are poorly aligned with dense contextual embeddings. Decision Tree also lagged behind the remaining classifiers, whereas KNN, LinearSVC, Logistic Regression, and Random Forest formed an intermediate group. This pattern reinforces the importance of reporting individual classifier results, since the cross-classifier average alone may hide substantial differences among learning algorithms.

5.2. Dataset-Level Robustness

Table 5 shows that aggregate scores obscure relevant corpus-level differences. Datasets such as FCN, FRECOGNA, FAKETRUEBR, and FAKEBR achieved high best-case macro-F_1 values,

Table 5. Dataset-level robustness on the test split. Mean macro-F_1 and performance range are computed across the 32 encoder–classifier pipelines.

Dataset	Test Samples	Mean Macro-F_1	Best Macro-F_1	Range	Best Pipeline
FCN	309	0.8532	0.9838	0.6512	albertina_ptbr_900m + svc_rbf
FRECOGNA	1786	0.7943	0.9266	0.5040	albertina_ptbr_900m + svc_rbf
FAKETRUEBR	537	0.7825	0.9552	0.7245	albertina_ptbr_900m + svc_rbf
FAKEBR	1080	0.7460	0.9491	0.6154	bertimbau_large + svc_rbf
FACTCKBR	316	0.7182	0.9319	0.5115	albertina_ptbr_100m + knn
TRE300	45	0.6830	0.9333	0.7204	bertimbau_large + svc_rbf
COVID19BR	360	0.6358	0.7973	0.2792	bertimbau_base + svc_rbf
MUMINPT	211	0.5179	0.6311	0.3358	albertina_ptbr_100m + random_forest
CENTRALFATOS	1547	0.5080	0.5817	0.0981	albertina_ptbr_900m + svc_rbf
FNEWSSET	90	0.4857	0.7129	0.3663	bertimbau_base + svc_rbf

whereas CENTRALFATOS, MUMINPT, and FNEWSSET remained more challenging on average. The range across pipelines also varied substantially, from 0.0981 in CENTRALFATOS to values above 0.7200 in FAKETRUEBR and TRE300, indicating different levels of sensitivity to the encoder–classifier choice.

These results reinforce the need for per-dataset analysis in a multi-corpus benchmark. A model may perform well in aggregate while still showing weaker behavior on specific corpora, especially under short texts, class imbalance, stylistic artifacts, or domain-specific patterns. They also explain why aggregate and dataset-balanced rankings may differ: aggregate metrics are influenced by corpus size, whereas the statistical analysis gives equal weight to each dataset. Thus, robustness should be assessed from both instance-level and corpus-level perspectives.

5.3. Statistical Significance Analysis

To assess whether the observed differences among the classical models were statistically reliable, we treated each dataset as the statistical unit and computed one macro-F_1 score for every dataset–model pair. The main analysis was performed on the test split, while the validation split was analyzed only as a complementary check. We first applied the Friedman test to compare all 32 encoder–classifier combinations jointly across the 10 complete datasets. We then selected the model with the highest mean dataset-level macro-F_1 and compared it against each remaining model using two-sided Wilcoxon signed-rank tests. The resulting p-values were adjusted with the Holm-Bonferroni correction to control the family-wise error rate. All statistical decisions used $\alpha = 0.05$.

Table 6. Statistical significance analysis over dataset-level macro-F_1 scores on the test split.

Statistic	Value
Number of models	32
Number of datasets	10
Friedman χ^2	161.9531
Friedman p-value	1.156×10^{-19}
Kendall's W	0.5224
Best mean model	bertimbau_large + svc_rbf
Best mean macro-F_1	0.8243
Significant post-hoc comparisons	0/31
Correction method	Holm-Bonferroni

The Friedman test indicated significant global differences among the 32 pipelines ($\chi^2 = 161.9531$, $p = 1.156 \times 10^{-19}$), with moderate agreement according to Kendall's $W = 0.5224$. The highest mean dataset-level macro-F_1 was achieved by `bertimbau_large + svc_rbf` (0.8243). However, none of the 31 pairwise post-hoc comparisons against this reference model remained statistically significant after Holm-Bonferroni correction. The validation split showed the same qualitative pattern, with a significant Friedman test but no significant corrected post-hoc comparison against the best average model.

These results provide an important qualification to the aggregate ranking. While there is clear global variation among model families, the small differences among the strongest pipelines should not be interpreted as definitive statistical superiority. In particular, the aggregate test table favors `albertina_ptbr_900m` under the best-configuration view, whereas the dataset-level average identifies `bertimbau_large + svc_rbf` as the highest mean model across corpora. This difference is expected because the aggregate view weights test instances directly, while the statistical analysis gives equal weight to each dataset. Taken together, the results suggest that the top contextual embeddings are highly competitive, and that classifier choice and corpus composition both influence the final ranking.

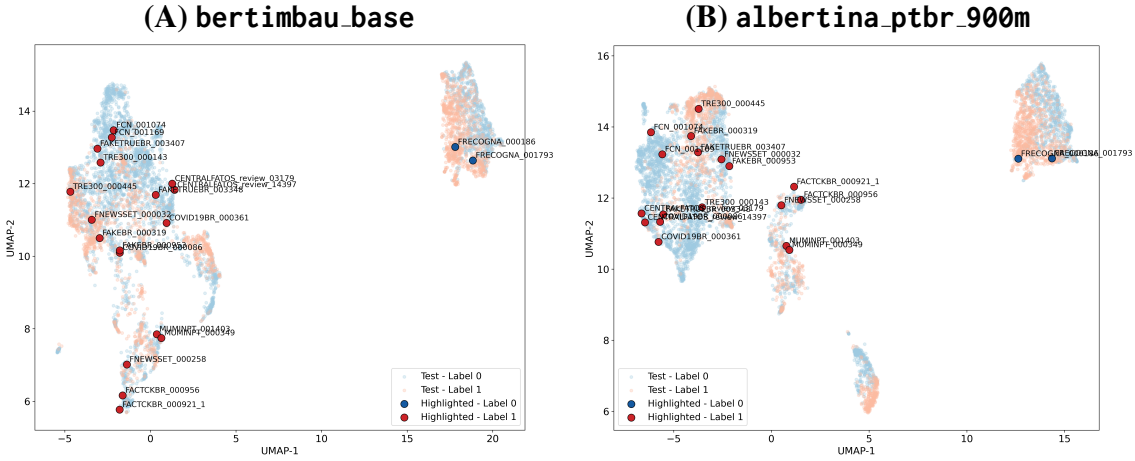

Figure 1. 2D UMAP projections of the balanced test set. (A) `bertimbau_base`; (B) `albertina_ptbr_900m`. Light colors denote the full test distribution by class; highlighted points indicate the most frequently misclassified samples across classical models.

The test-set UMAP projections in Figure 1 provide a qualitative complement to the dataset-level results. In both embeddings, the data form well-defined manifolds, but the structure appears to be influenced more by corpus-specific geometry than by the binary label alone. This pattern is consistent with Table 5, where different corpora exhibit distinct average macro-F_1 values and ranges across pipelines.

The highlighted points, corresponding to the most frequently misclassified instances, tend to appear near boundary regions or sparse neighborhoods. This suggests that some errors are associated with intrinsically difficult documents whose surface form resembles the opposite class. In particular, several hard cases are *REAL* news items containing lexical or rhetorical cues commonly associated with misleading content, such as

controversial claims, politically charged entities, or polarized discourse.

Overall, the results show that dense Portuguese contextual embeddings combined with classical classifiers can achieve strong performance in multi-corpus fake news classification. The aggregate benchmark favors `albertina_ptbr_900m` in terms of peak test performance and average performance across classifiers, while the dataset-level statistical analysis identifies `bertimbau_large + svc_rbf` as the highest mean macro-F_1 configuration across datasets. However, the corrected post-hoc analysis does not support definitive statistical superiority among the strongest configurations. Thus, the central finding should be interpreted as a joint effect of embedding quality, classifier choice, and dataset heterogeneity. In particular, `svc_rbf` was the most consistent downstream classifier, larger contextual encoders tended to provide stronger representations, and the per-dataset analysis revealed that some corpora remain substantially more difficult than others. This broader view strengthens the empirical contribution of the benchmark by moving beyond a single aggregate ranking and characterizing how encoder–classifier combinations behave under heterogeneous Portuguese fake news corpora.

6. Final considerations and conclusion

This study presented a multi-corpus benchmark for fake news classification in Brazilian Portuguese, combining four contextual embeddings with eight classical classifiers under a unified protocol. Dense Portuguese-specific embeddings paired with classical classifiers achieved strong performance: `albertina_ptbr_900m` obtained the best peak and the highest average macro-F_1 across classifiers, while `bertimbau_large` achieved the strongest validation results and the highest dataset-level mean macro-F_1 when paired with `SVC-RBF`. Still, the statistical analysis indicates that small differences among the top configurations should be interpreted with caution.

Performance depends on the interaction between embedding, classifier, and dataset. `SVC-RBF` was the most consistent classifier, and simpler models such as Logistic Regression and Linear SVM also benefited from richer embedding spaces. The per-dataset analysis revealed substantial heterogeneity across corpora, suggesting that aggregate scores may hide variations in robustness, while the qualitative inspection of misclassified samples showed that the hardest cases were predominantly *REAL* items whose wording or thematic framing resembled misleading content.

From an applied standpoint, the frozen-embedding plus classical-classifier pipeline offers practical advantages over end-to-end fine-tuned transformers: lower training and inference cost, reduced memory footprint, and greater interpretability via feature-importance and decision-boundary analyses. These properties make the approach particularly suitable for deployment in low-resource or auditable settings, supporting public-interest misinformation-detection tools where transparency, reproducibility, and operational efficiency are as relevant as raw predictive performance.

Future work should extend this benchmark with recent Portuguese and multilingual encoders, sentence-level representations, and embeddings trained for misinformation-related tasks; compare fine-tuned transformers and parameter-efficient adaptation against the frozen-embedding classical setting adopted here; and incorporate datasets with different domains, text lengths, temporal periods, and annotation protocols to assess whether the observed patterns hold under broader scenarios.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Cantarino, F. H. S. (2024). Criação de um corpus português para auxiliar a identificação de notícias verdadeiras e falsas. Trabalho de Conclusão de Curso, Universidade Federal de Uberlândia. Corpus provenance for BoatosBr.

Chavarro, J. P., Carvalho, J. T., Portela, T. T., and Silva, J. C. (2023). Faketruebr: Um corpus brasileiro de notícias falsas. In *Anais da Escola Regional de Banco de Dados*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

da Silva, F. R. M., Freire, P. M. S., de Souza, M. P., de A. B. Plenamente, G., and Goldschmidt, R. R. (2020). Fakenewssetgen: A process to build datasets that support comparison among fake news detection methods. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

de Morais, J. I., Abonizio, H. Q., Tavares, G. M., da Fonseca, A. A., and Barbon Jr., S. (2020). A multi-label classification system to distinguish among fake, satirical, objective and legitimate news in brazilian portuguese. *iSys – Brazilian Journal of Information Systems*, 13(4):126–149.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Garcia, G. L., Afonso, L. C. S., and Papa, J. P. (2022). Fakerecogna: A new brazilian corpus for fake news detection. In *International Conference on Computational Processing of the Portuguese Language*, pages 57–67.

Gôlo, M. P. S., Mori, A. L. V., Oliveira, W. G., Barbosa, J. R., Graciano-Neto, V. V., Lima, E. A. d., and Marcacini, R. M. (2024). On the use of large language models to detect brazilian politics fake news. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional*.

Gôlo, M. P. S., Souza, M. C. d., Rossi, R. G., Rezende, S. O., Nogueira, B. M., and Marcacini, R. M. (2023). One-class learning for fake news detection through multi-modal variational autoencoders. *Engineering Applications of Artificial Intelligence*, 124:106088.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley.

Jeronimo, C. L. M., Marinho, L. B., Campelo, C. E. C., Veloso, A., and Melo, A. S. d. C. (2019). Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, pages 15–24.

Marques, I., Salles, I., Couto, J. M. M., Pimenta, B. C., Assis, S., Reis, J. C. S., da Silva, A. P. C., Almeida, J. M., and Benevenuto, F. (2022). A comprehensive dataset of brazilian fact-checking stories. *Journal of Information and Data Management*, 13(1).

Martins, A. D. F., Cabral, L., Mourão, P. J. C., de Sá, I. C., Monteiro, J. M., and Machado, J. (2021). Covid19.br: A dataset of misinformation about covid-19 in brazilian portuguese whatsapp messages. In *Anais do Dataset Showcase Workshop*.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., 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*, pages 324–334.

Moreno, J. G. and Bressan, G. C. (2019). Factck.br: A new dataset for claim detection and related fact checking. In *Proceedings of the International Conference on the Computational Processing of Portuguese*.

Nielsen, D. S. and McConville, R. (2022). Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153.

Paixão, M., Lima, R., and Espinasse, B. (2020). Fake news classification and topic modeling in brazilian portuguese. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Pires, V. B. and Guerreiro e Silva, D. (2024). Portuguese fake news classification with bert models. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional*, pages 834–845.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*.

Santos, R., Rodrigues, J., Gomes, L., Silva, J., Branco, A., Cardoso, H. L., Osório, T. F., and Leite, B. (2024). Fostering the ecosystem of open neural encoders for portuguese with albertina pt-* family.

Santos, R. L. d. S. (2022). *Detecção automática de notícias falsas em português*. Tese de doutorado, Universidade de São Paulo. Acesso em: 30 mar. 2026.

Silva, R. M., Santos, R. L. S., Almeida, T. A., and Pardo, T. A. S. (2020). Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113199.

Sousa, F., Barbosa, A., Oliveira, C., and Braga, R. (2022). Detecção de fake news em língua portuguesa combinando redes neurais convolucionais e algoritmos de aprendizagem de máquina. In *Anais do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 336–348.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*.

Villela, H. F., Corrêa, F., Ribeiro, J. S. d. A. N., Rabelo, A., and Carvalho, D. B. F. (2023). Fake news detection: a systematic literature review of machine learning algorithms and datasets. *Journal on Interactive Systems*, 14(1).

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Zhang, H. (2004). The optimality of naive bayes. *AA*, 1(2):3.