Language Models and Detection Strategies for Political Disinformation in Brazil: A Comparative Applied Study

Adriel L. V. Mori[1,2], Willgnner Ferreira Santos[1], Eliomar Araújo Lima[1], Valdemar V. Graciano Neto[1], Jacson R. Barbosa[1], Rogerio R. Carvalho[1], Nádia F. F. da Silva[1] 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,willgnner_santos}@discente.ufg.br

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

Abstract. *The rapid spread of online political disinformation has increased the need for scalable detection beyond manual fact-checking. This study evaluates textual disinformation detection in Brazilian Portuguese by comparing BERTimbau-based supervised models with open-source LLMs under zero-shot, Chain-of-Thought, and RAG settings. Four public corpora were used for training, validation, and retrieval, while an external political dataset supported out-of-distribution testing. Results show that gemma2:9b in zero-shot achieved the best macro-F_1 (0.950), outperforming BERTimbau+SVM (0.837). BM25-based RAG reached macro-F_1 = 0.923 with grounded decisions. Overall, performance depends on inference strategy, architecture, and corpus characteristics.*

1. Introduction

Political disinformation circulating through digital platforms has become a central element of Brazil's communication ecosystem, driven by misleading content that spreads faster and more broadly than verified news [Vosoughi et al. 2018]. This asymmetry becomes particularly critical during electoral periods and public crises, such as the 2018 Brazilian elections, marked by fraud rumors and intense misinformation flows, and the Coronavirus Disease 2019 (COVID-19) pandemic, when the so-called "infodemic" amplified the circulation of misleading health-related content [Bernardi 2021]. In response, Brazil has advanced both regulatory initiatives, such as the Marco Civil da Internet [Schreiber 2022] and electoral resolutions issued by the Superior Electoral Court (TSE) [BRASIL 2019], and technical efforts based on language models for automated detection. Even so, substantial challenges remain regarding accuracy, interpretability, and robustness under real-world conditions.

Despite this progress, research on political disinformation detection in Brazilian Portuguese remains methodologically fragmented: existing studies typically rely on multilingual corpora [Nielsen and McConville 2022], a single dataset [Pires et al. 2024], or restricted evaluation settings that do not account for domain shift, and no prior work offers an integrated comparison combining multiple Portuguese-language corpora, open-source LLMs, and heterogeneous inference regimes under out-of-distribution conditions in the political domain. Consequently, it remains empirically unresolved which combinations of

model family, inference strategy, and data configuration are most effective for handling the thematic, temporal, and stylistic variability of Brazilian fake news [Silva et al. 2025]. To address this gap, this article evaluates a broad set of approaches, spanning supervised classifiers built on BERTimbau embeddings, a fine-tuned BERTimbau model, and open-source LLMs under zero-shot, Chain-of-Thought (CoT), and Retrieval-Augmented Generation (RAG) regimes, under a unified OOD protocol in which all paradigms share the same training and validation corpora, the same external political test set, and the same evaluation metrics, enabling controlled and reproducible cross-paradigm comparisons.

The results show that open-source LLMs are competitive with supervised approaches. Fine-tuned BERTimbau obtained the strongest supervised result, whereas the highest overall performance was achieved by `gemma2:9b` in the zero-shot setting (macro-$F_1 = 0.950$). RAG-based variants exhibited a more robust error profile across thematically diverse cases, though without surpassing the best zero-shot result. The proposed pipeline relies exclusively on open-source models and is publicly available to support reproducibility.

The main contributions of this work to Natural Language Processing (NLP) in Portuguese are as follows:

- **Unified OOD benchmark.** We consolidate a multi-corpus benchmark for political disinformation in Brazilian Portuguese and establish a rigorous evaluation protocol in which a held-out external political test set is strictly isolated from training data, retrieval indices, and prompt design, preventing information leakage across paradigms.
- **Cross-paradigm comparative protocol.** We design and apply a controlled evaluation framework that places supervised classifiers, fine-tuned transformers, and open-source LLMs under zero-shot, CoT, and RAG regimes on identical experimental conditions, enabling direct cross-paradigm comparisons, an aspect underexplored in existing Portuguese-language work.
- **Qualitative error analysis via LLM-generated justifications.** We examine model-generated explanations to characterize systematic failure modes under distribution shift, identifying recurring patterns such as the projection of causal hypotheses not grounded in the source text, complementing aggregate metrics and informing practical deployment.
- **Open-source reproducible framework.** We release all code, preprocessing pipelines, and experimental configurations to support reproducibility and future research on political disinformation detection in Portuguese.

2. Related work

Recent literature on disinformation detection in Portuguese relies on a relatively consolidated set of corpora, summarized in Table 1. Although this does not constitute a formal systematic review, the bibliographic survey conducted points to a heterogeneous ecosystem that includes pioneering resources such as Fake.Br [Monteiro et al. 2018], datasets derived from journalistic fact-checking initiatives [Moreno and Bressan 2019, Couto et al. 2021], data collected from social media and messaging platforms [Cordeiro and Pinheiro 2019, Cabral et al. 2021], as well as more recent collections focused on political disinformation and general news [Chavarro et al. 2023, Gôlo et al. 2024, Garcia et al. 2024]. This landscape demonstrates significant progress in the availability of Portuguese-language data,

yet it also reveals the predominance of binary or strictly multiclass classification tasks and the scarcity of native corpora explicitly designed for automated fact verification. This gap has motivated initiatives based on the translation and adaptation of well-established English-language resources, such as LIAR and AVeriTeC [Delucis et al. 2025].

Year	Corpus	Domain	Multilingual	Multilabel	Public	Data Period
2018	Fake.Br [Monteiro et al. 2018]	Online news (general)			✓	2016–2017
2019	FactCK.Br [Moreno and Bressan 2019]	Journalistic fact-checking			✓	2014–2019
2019	FakeTweet.Br [Cordeiro and Pinheiro 2019]	Twitter rumors			✓	2016–2019
2019	BRACIS One-Class [Faustini and Covões 2019]	General fake news				2018–2019
2020	De Morais et al. [de Morais et al. 2020]	News (mixed)		✓		2010–2020
2020	FakeNewsSetGen [Silva 2020]	Multidomain (synthetic generation)			✓	Various
2020	MM-COVID [Li et al. 2020]	COVID-19	✓		✓	2020
2020	FakeCovid [Shahi and Nandini 2020]	COVID-19	✓		✓	2020
2021	FakeWhatsApp.Br [Cabral et al. 2021]	WhatsApp messages			✓	2018–2021
2021	COVID19.Br [Martins et al. 2021]	COVID-19 on WhatsApp			✓	2020–2021
2021	FCN (Fact Checked News) [Gôlo et al. 2021]	Verified multi-domain news			✓	Various
2021	Central de Fatos [Couto et al. 2021]	Fact-check aggregation platform			✓	2013–2021
2022	MuMiN-PT [Nielsen and McConville 2022]	Social media fact-checking	✓		✓	2020–2022
2022	FakeRecogna [Garcia et al. 2022]	General news (PT-BR)			✓	2017–2021
2022	Fakepedia [Charles et al. 2022]	General news		✓	✓	Various
2023	FakeTrue.Br [Chavarro et al. 2023]	True/false news			✓	2018–2023
2024	Boatos.Br [Cantarino 2024]	General news and rumors			✓	Various
2024	LLM Fake News Br [Gôlo et al. 2024]	Politics (PT-BR)			✓	2022–2024
2024	WikiNews-PT [Garcia et al. 2024]	News classification			✓	2022
2025	WhatsApp Political Groups BR [Kansaon et al. 2025]	Political coordination on WhatsApp			✓	Jul. 2022–2023

Table 1. Main corpora used in disinformation detection and news classification tasks.

From a methodological perspective, research in Portuguese spans classical approaches based on TF–IDF vectors and stylometric features [Cantarino 2024] to pretrained Transformer architectures such as BERT, mBERT, and BERTimbau [Pires et al. 2024], with evidence that domain-specialized models can be competitive with multilingual alternatives in Portuguese NLP tasks [Herculano et al. 2025]. In parallel, the English-language literature explores LLMs on corpora such as LIAR [Boumber et al. 2024], hybrid RAG systems [Lopez-Joya et al. 2025], and CoT and in-context learning strategies [Hu et al. 2024], while recent work highlights that LLM performance in content classification is sensitive to prompt design and event-specific context [Paiva et al. 2025]. Despite these advances, studies remain fragmented and rely on a single corpus or on proprietary models [Delucis et al. 2025, Kansaon et al. 2025], and LLMs and generative approaches, though increasingly competitive with supervised methods, still face challenges under emerging linguistic patterns [Silva et al. 2025]. This article addresses these limitations through a unified comparative evaluation of open-source LLMs and supervised baselines under the same OOD protocol across multiple Portuguese-language corpora.

3. Methodology

The overall pipeline comprises: (i) the use of aggregated Portuguese-language corpora as experimental resources for supervised training and retrieval, (ii) evaluation on an independent dataset focused on contemporary Brazilian politics, and (iii) comparison across three modeling settings, namely zero-shot LLMs, supervised BERTimbau-based models, and RAG-based architectures.

3.1. Datasets

Five corpora were used in this study, as summarized in Table 2: FakeRecogna, FakeTrueBr, BoatosBr, MuMiN-PT, and LLM Fake News Br. The first four were aggregated to form

the training and validation set, covering different circulation contexts such as web pages, WhatsApp, Facebook, and Twitter/X, and labeled under a binary veracity scheme (REAL vs. FAKE) based on fact-checking procedures. These corpora were selected because they are the most compatible with the external test set, LLM Fake News Br, in terms of language, task formulation, and news verification setting. Although Table 1 presents a broader set of Portuguese-language resources, most were excluded because they are multilingual, multilabel, synthetic, non-public, insufficiently textual, or not directly compatible with the binary veracity setting adopted here. The final selection therefore prioritizes task compatibility, public availability, source diversity, and reproducibility, while also helping reduce source bias by covering distinct communication channels.

Table 2. Class distribution across the corpora used in this study.

	Datasets				
Class	**BoatosBr**	**FakeTrueBr**	**FakeRecogna**	**MuMiN-PT**	**LLM Fake News Br**
FAKE	1888	1791	5951	1339	148
REAL	1516	1791	5951	65	152
Total	3404	3582	11902	1404	300

The LLM Fake News Br corpus is reserved as the external test set. As shown in Table 2, it contains 300 approximately balanced instances and focuses on recent political and electoral content, including material associated with the 2018 and 2022 elections and electoral fact-checking activities linked to the TSE. This choice allows the evaluation to target a sensitive domain under an out-of-distribution setting.

The FakeRecogna, FakeTrueBr, BoatosBr, and MuMiN-PT corpora were used as the aggregated training and validation set for the supervised experiments, including both classical models operating on BERTimbau embeddings and end-to-end fine-tuned architectures, as well as for RAG retrieval when applicable. In contrast, zero-shot LLMs were not adapted to these data and were evaluated directly on LLM Fake News Br, which was kept separate from both the supervised training data and the retrieval index. This design prevents information leakage and enables a proper out-of-distribution evaluation. Across all datasets considered in this study, the class distribution shows a slight imbalance toward the FAKE class, which represents approximately 54% of the samples, whereas the REAL class accounts for about 46%.

3.2. Inference models and strategies

The evaluated approaches are organized into LLMs operating without supervised training and supervised models. Under the zero-shot regime, the models are executed locally via Ollama, following the protocol described in [Gôlo et al. 2024]. Each news item is presented through a direct binary classification prompt, in which the model is instructed to assign only one of two possible labels, `FAKE` or `REAL`. This formulation enforces a standardized and directly comparable decision process across architectures. For all LLM-based experiments, the decoding parameters were fixed at `temperature=0`, `top_k=1`, and `top_p=1` to enforce greedy decoding and maximize output determinism. It should be noted that, even under these settings, minor non-deterministic variation may occur when running on parallel GPU hardware; however, under the single-GPU local execution adopted in this

study via Ollama, outputs are expected to be fully reproducible across runs. A simplified version of the prompt is shown below:

> You are a binary news classifier.
> Task: determine whether the following text is FAKE or REAL.

For interpretability purposes, a second and independent inference is performed with a dedicated prompt for brief justification generation, without affecting the final classification output.

As an extension of this setting, a zero-shot CoT prompting is also evaluated. In this case, the model is instructed to reason through a fixed sequence of verification steps before producing a final standardized label. The reasoning structure explicitly covers factual plausibility, sentiment-related manipulation cues, and logical consistency. This design remains fully zero-shot while introducing an interpretable reasoning scaffold during inference, broadly aligned with recent approaches that employ structured intermediate reasoning in fact-checking [Liu et al. 2025]. A simplified fragment of the prompt is shown below:

> You are a fact-checking assistant. Classify the text as FAKE or REAL.
> Think step by step and, at the end, provide only one final standardized line.
> 1. Fact verification: assess whether the claims are plausible or contradictory.
> 2. Sentiment analysis: identify alarmist, emotional, or manipulative language.
> 3. Logical consistency: verify internal contradictions or lack of evidence.

This strategy allows the analysis of whether explicit reasoning improves decision consistency and classification performance without introducing supervised adaptation.

Table 3. Language models evaluated in the zero-shot setting.

Developer	Version	Parameters	Quantization	Size (GB)
Google DeepMind	gemma3	12.2B	Q4_K_M	7.59
Meta	llama3.1	8.0B	Q4_K_M	4.58
Meta	llama3.2	3.2B	Q4_K_M	1.88
Microsoft	phi3	14.0B	Q4_0	7.35
Alibaba	qwen2.5	14.8B	Q4_K_M	8.37
Alibaba	qwen2.5	7.6B	Q4_K_M	4.36
Google DeepMind	gemma2	9.2B	Q4_0	5.07
Alibaba	qwen2	7.6B	Q4_0	4.13

Within the supervised setting, BERTimbau was evaluated under two configurations. First, it was used as a fixed encoder to generate dense textual embeddings, which were subsequently provided as input to classical classifiers, including Logistic Regression, Linear SVM, RBF-SVM, Random Forest, Gradient Boosting, and Gaussian Naive Bayes [Patel et al. 2022]. The train and validation subsets were defined through a stratified procedure that preserved the joint distribution of dataset origin and class label, while model selection for the classical classifiers was carried out by grid search with stratified cross-validation. Second, BERTimbau was evaluated in an end-to-end fine-tuning configuration, in which both the Transformer encoder and the final classification layer were updated during supervised learning. This experimental design enables a direct comparison between

classical classifiers operating on fixed contextual representations and full supervised adaptation of the language model under the same train-validation-test protocol.

Table 3 presents the language models selected for zero-shot analysis, including version, parameter count, quantization scheme, and disk size. All models are fully open-source and were executed locally without relying on any proprietary API, ensuring reproducibility and eliminating access barriers. Although a formal analysis of inference cost per sample is left for future work, model size and quantization serve as proxies for computational demand: larger models such as `phi3` (14.0B, 7.35 GB) and `qwen2.5:14b` (14.8B, 8.37 GB) impose substantially higher memory and latency requirements than smaller alternatives such as `llama3.2` (3.2B, 1.88 GB), with only modest gains in macro-F_1, suggesting that mid-sized models in the 7–9B range offer a favorable performance-to-cost trade-off for practical deployment.

3.3. RAG architectures and re-ranking

The third group comprises Retrieval-Augmented Generation (RAG) architectures. In this setting, the classification decision is based on external evidence retrieved from a knowledge base. This knowledge base is built only from labeled instances in the training and validation corpora (FakeRecogna, FakeTrueBr, BoatosBr, and MuMiN-PT). Each instance is converted into a textual passage using the title and body when available, or the main textual content otherwise. The external test set (LLM Fake News Br) is not included in the retrieval index. This choice prevents information leakage and preserves the out-of-distribution evaluation protocol.

Given a target news item, the retrieval stage first builds a candidate pool of relevant passages using two strategies: lexical retrieval with BM25 and dense retrieval with FAISS over embedding representations. From this pool, a small set of evidence passages, usually with $k = 5$, is selected and inserted into the prompt together with the target news item. In this design, the aggregated corpus serves as a retrieval source rather than as an optimization resource for zero-shot LLMs. We also evaluate a re-ranking variant, in which a broader candidate set is first retrieved and then reordered according to semantic relevance to the target item before prompt construction. Unless otherwise stated, the main results reported in the paper correspond to the best-performing retrieval configuration.

3.4. Evaluation protocol

The task is formulated as binary classification, with the positive class corresponding to fake news. FakeRecogna, FakeTrueBr, BoatosBr, and MuMiN-PT are used for training and validation, while LLM Fake News Br is reserved exclusively as the external test set, ensuring strict out-of-distribution evaluation. Macro-F_1 and FAKE-class recall are adopted as primary metrics, as they jointly capture overall discriminative performance and sensitivity to disinformative content.

Within the supervised setting, two configurations are evaluated under the same train-validation-test split: classical classifiers trained over fixed BERTimbau embeddings, covering Logistic Regression, Linear SVM, RBF-SVM, Random Forest, Gradient Boosting, and Gaussian Naive Bayes; and end-to-end fine-tuned BERTimbau, in which both the encoder and the classification layer are updated during training. This design enables a principled contrast between feature-based and fully adaptive supervised learning under identical evaluation conditions.

All experiments are conducted under a homogeneous protocol with deterministic decoding settings. A single prompt per inference strategy is applied consistently across all models to minimize prompt-related variability and ensure comparability. Repeated executions with standard deviation reporting are left for future work. All code, configurations, and corpus preprocessing instructions are publicly available, subject to their original licenses, in the project repository[1].

4. Results

The experiments followed the protocol described in Section 3, using the LLM Fake News Br dataset as the test set, with binary labels (FAKE vs. REAL) on recent political news. For all approaches, we report class-wise F_1 scores and macro-F_1, with particular emphasis on the FAKE class, given the high cost of false negatives in disinformation settings.

4.1. LLMs under a zero-shot regime

As reported in Table 4, `gemma2:9b` achieves the best overall result, with a macro-F_1 of 0.950 and well-balanced scores across the FAKE ($F_1 = 0.948$) and REAL ($F_1 = 0.952$) classes. This behavior indicates strong generalization and an appropriate trade-off between sensitivity to disinformation and the preservation of true news, which is desirable in operational settings.

Table 4. Per-class metrics and macro-average F_1 for veracity classification (`label`) across all LLMs in the zero-shot setting.

Model	FAKE			REAL			Macro-F_1
	Precision	Recall	F_1	Precision	Recall	F_1	
`gemma2:9b`	0.972	0.926	0.948	0.931	0.974	0.952	0.950
`qwen2.5:14b`	1.000	0.750	0.857	0.804	1.000	0.891	0.874
`gemma3:12b`	0.966	0.764	0.853	0.809	0.974	0.884	0.868
`llama3.1:8b`	0.963	0.701	0.811	0.771	0.974	0.860	0.836
`qwen2.5:7b`	0.919	0.615	0.737	0.716	0.947	0.816	0.776
`qwen2:7b`	0.706	0.811	0.755	0.785	0.671	0.723	0.739
`phi3:14b`	0.930	0.446	0.603	0.642	0.967	0.772	0.687
`llama3.2:3b`	0.921	0.392	0.550	0.620	0.967	0.756	0.653

The introduction of CoT prompting changes the performance profile of the models. As shown in Table 5, the best CoT results are again achieved by `gemma2:9b` and `gemma3:12b`, with macro-F_1 around 0.907–0.908. Although these values are lower than those obtained under the instructional zero-shot setting, we observe greater homogeneity among mid-sized models and less extreme variation across classes, suggesting more stable decisions.

The CoT approach does not yield absolute performance gains relative to the instructional zero-shot setting, but it promotes greater consistency across different architectures. This stability suggests that explicit reasoning acts as an inference regularization mechanism, reducing sharp fluctuations and making model behavior more predictable, albeit at the cost of a slight decrease in overall metrics.

[1] `https://github.com/adrielmori/sbbd2026-249221`

Table 5. Per-class metrics and macro-average F_1 for veracity classification (*label*) across all LLMs under the zero-shot CoT approach.

Model	FAKE			REAL			Macro-F_1
	Precision	Recall	F_1	Precision	Recall	F_1	
gemma2:9b	0.890	0.926	0.907	0.931	0.888	0.909	0.908
gemma3:12b	0.885	0.932	0.908	0.931	0.882	0.905	0.907
llama3.1:8b	0.877	0.723	0.793	0.834	0.895	0.863	0.828
llama3.2:3b	0.792	0.824	0.808	0.838	0.783	0.810	0.809
phi3:14b	0.904	0.696	0.786	0.758	0.928	0.834	0.810
qwen2.5:14b	0.933	0.845	0.887	0.861	0.941	0.899	0.893
qwen2.5:7b	0.887	0.845	0.865	0.855	0.895	0.875	0.870
qwen2:7b	0.878	0.730	0.797	0.774	0.901	0.833	0.815

4.2. RAG architectures

RAG-based architectures produced more stable decisions than pure zero-shot inference, although they did not surpass the best instructional LLM. The best individual result was obtained by gemma2:9b with BM25 and no re-ranking, reaching macro-$F_1 = 0.923$, with $F_1 = 0.921$ for FAKE and $F_1 = 0.926$ for REAL. This result indicates a balanced behavior across classes when the decision is supported by retrieved evidence.

Table 6. Macro-F_1 for RAG configurations.

Model	BM25	BM25 + rerank	FAISS	FAISS + rerank
gemma2:9b	**0.923**	0.903	0.910	0.903
gemma3:12b	0.845	**0.862**	0.839	0.818
llama3.1:8b	0.720	0.690	**0.743**	0.693
llama3.2:3b	**0.398**	0.378	**0.398**	0.391
phi3:14b	**0.745**	0.729	0.673	0.731
qwen2.5:14b	0.828	0.810	**0.853**	0.815
qwen2.5:7b	0.428	0.413	**0.687**	0.569
qwen2:7b	0.779	0.772	0.765	**0.783**
Mean macro-F_1	0.708	0.695	**0.733**	0.713
Best individual result	**0.923**	0.903	0.910	0.903

Table 6 shows that re-ranking did not produce consistent gains across models or retrieval strategies. In several cases, the version without re-ranking achieved equal or better results. For this reason, the main discussion in this article focuses on the best-performing configuration, BM25 without re-ranking, while the full set of RAG variants is reported in the repository.

4.3. Supervised models with fixed BERTimbau embeddings

Among the classical classifiers built on fixed BERTimbau embeddings, the best validation result was obtained by the RBF-SVM, which reached macro-$F_1 = 0.945$. The second-best result was achieved by Gradient Boosting, with macro-$F_1 = 0.918$, followed by Linear SVM and Logistic Regression, both with macro-F_1 above 0.90. Random Forest also showed competitive validation performance, with macro-$F_1 = 0.901$, whereas Gaussian Naive Bayes obtained the lowest result in this setting, with macro-$F_1 = 0.801$.

On the external test set, performance decreased for all classical models, confirming the difficulty of out-of-distribution evaluation. The best result was again obtained by the RBF-SVM, with macro-$F_1 = 0.837$, followed very closely by Random Forest, with macro-$F_1 = 0.836$. Gradient Boosting reached macro-$F_1 = 0.801$, while Linear SVM and Logistic Regression remained below 0.79. Gaussian Naive Bayes had the weakest test result, with macro-$F_1 = 0.508$. Overall, the highest macro-F_1 on the external test set among the classical models was achieved by the RBF-SVM.

4.4. Fine-tuned BERTimbau

BERTimbau was also evaluated in an end-to-end fine-tuned setting for binary classification. The model was trained with the `neuralmind/bert-base-portuguese-cased` checkpoint, `max_length=256`, learning rate of 2×10^{-5}, batch sizes of 8 for training and 16 for evaluation, and 3 epochs. In this setting, both the Transformer encoder and the final classification layer were updated during training.

The fine-tuned model achieved strong performance on both validation and test. On the validation split, it reached macro-$F_1 = 0.974$, indicating stable supervised learning on the aggregated corpus. On the external test set, it achieved macro-$F_1 = 0.916$, with recall equal to 1.000 for the FAKE class. This result is particularly relevant for disinformation detection, since it shows that all fake instances in the test set were correctly identified.

5. Conclusion

Table 7. Summary of the best results obtained for each evaluated approach. Class-wise metrics and the macro-F_1 score are reported.

Approach	*FAKE*			*REAL*			Macro-F_1
	Precision	**Recall**	F_1	**Precision**	**Recall**	F_1	
Zero-shot (`gemma2:9b`)	0.972	0.926	0.948	0.931	0.974	0.952	**0.950**
Zero-shot + CoT (`gemma2:9b`)	0.890	0.926	0.907	0.931	0.888	0.909	0.908
RAG (BM25, without reranking)	0.937	0.905	0.921	0.911	0.941	0.926	0.923
BERTimbau + SVM	0.760	0.986	0.859	0.981	0.697	0.815	0.837
BERTimbau Fine-tuned	0.855	1.000	0.922	1.000	0.836	0.910	0.916

Table 7 summarizes the best results across all evaluated paradigms. The zero-shot configuration with `gemma2:9b` achieved the strongest overall performance, with the highest accuracy (0.950), macro-F_1 (0.950), and MCC (0.901), committing only 15 errors in the test set. Its class-wise performance was the most balanced (FAKE $F_1 = 0.948$, REAL $F_1 = 0.952$), and remains strong under bootstrap analysis, with a 95% confidence interval of $[0.923,\ 0.973]$ for macro-F_1. These numbers indicate that instruction-tuned LLMs can capture relevant semantic and contextual cues without task-specific supervision while maintaining a stable decision profile across classes.

Among the supervised settings, fine-tuned BERTimbau obtained the best result (accuracy $= 0.917$, macro-$F_1 = 0.916$, MCC $= 0.845$), clearly outperforming the fixed-embedding SVM baseline (macro-$F_1 = 0.837$, MCC $= 0.712$), a difference confirmed by pairwise McNemar analysis. CoT prompting did not improve the best overall score but shifted the decision profile, increasing sensitivity to FAKE at the cost of more false positives, suggesting it modifies the model's operating point rather than peak performance.

More broadly, the statistical evidence shows that the strongest methods consistently outperform the SVM baseline, whereas differences among the top-performing models are smaller and less clear-cut.

RAG-based architectures did not reach the top score, but they remain relevant in the comparison because they changed the distribution of errors and provided a more evidence-grounded decision process. In the reported setting, RAG achieved macro-$F_1 = 0.873$ and MCC$= 0.747$, outperforming the SVM baseline while remaining below zero-shot, CoT, and fine-tuned BERTimbau. At the same time, the broader comparison shows that most test instances were easy for all methods, with 209 of 300 items correctly classified by all five models and 263 of 300 correctly classified by at least four models. This suggests that the main differences between approaches are concentrated in a relatively small but difficult subset of politically sensitive cases. Overall, the results reinforce that inference strategy, distribution shift, and error profile are central factors for political disinformation detection in Portuguese, and that the best-performing approaches should be interpreted not only by macro-F_1, but also by class balance, MCC, and pairwise statistical comparison.

The error-oriented analysis revealed that the instances with the highest error rates are concentrated in the REAL class. A representative example is the news item `TRE300_000238`, which exhibited an error rate above 0.71 and was incorrectly classified by multiple approaches in at least one inference run.

> "Who is the most powerful person in Brazil today? There is unlikely to be a consensus, but the names of Lula (PT), the Speaker of the Chamber of Deputies Arthur Lira (PP-AL), and Supreme Federal Court (STF) Justice Alexandre de Moraes would likely appear among the answers... three authorities who have defended, for more than a year and without success, the need to pass regulation for digital platforms... Elon Musk's move against Moraes is merely the most flamboyant episode of big tech interference in Brazilian affairs."

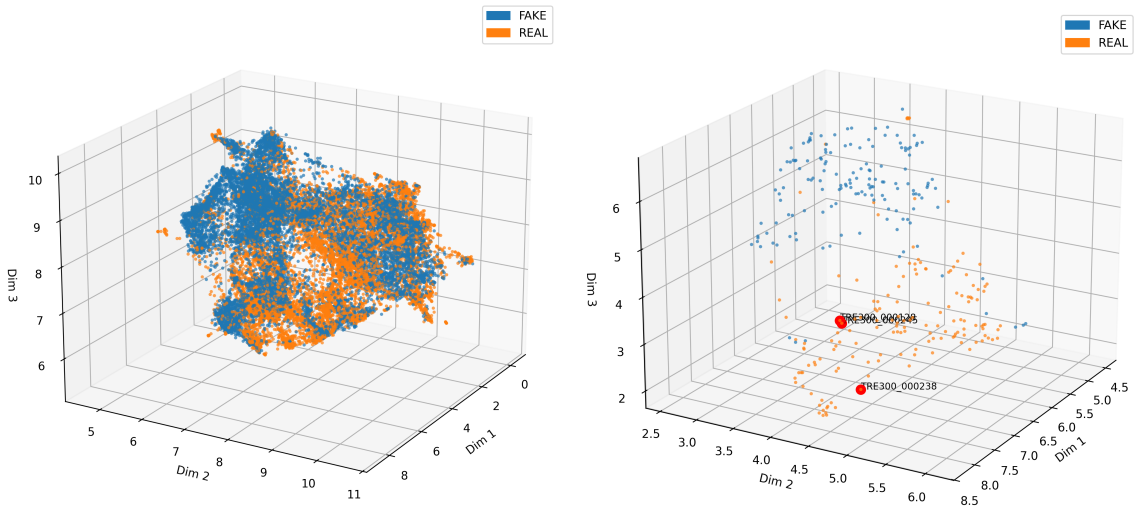
Although the content is analytical, contextualized, and factually consistent, its argumentative density and the use of politically sensitive references make this instance particularly challenging. This difficulty is reflected in the divergent predictions produced by the evaluated models.

When analyzed under the RAG architecture with BM25 retrieval (no re-ranking), `gemma2:9b` produced a justification aligned with the text and correctly preserved the REAL label:

> "The news item is considered true because it presents evidence and concrete facts about the influence of big tech companies in Brazil and the difficulty of passing the so-called 'Fake News Bill'."

In contrast, under the same retrieval configuration, llama3.1:8b produced an incorrect justification, introducing inferences that are not supported by the text:

> "The news item may be classified as false because there is no concrete evidence that Elon Musk's statements were coordinated with a far-right campaign against the bill."



(a) Three-dimensional (3D) UMAP projection of the training embeddings.

(b) 3D UMAP projection highlighting instances with the highest error rates.

Figure 1. Three-dimensional UMAP projections of the textual embeddings: (a) training set distribution and (b) instances with the highest error frequency across models and inference strategies.

This example illustrates a recurring error pattern: the tendency of some LLMs to project external causal hypotheses or implicit narratives that are not explicitly present in the original text, leading to false positives. The side-by-side comparison of the responses highlights the role of evidence retrieval and textual grounding as central factors for more reliable decisions.

6. Final considerations, limitations, and ethical implications

This study advances political disinformation detection in Brazilian Portuguese by consolidating a multi-corpus benchmark and evaluating supervised baselines and open-source LLMs under a unified out-of-distribution protocol. `gemma2:9b` achieved the best overall performance in the zero-shot setting (macro-$F_1 = 0.950$, MCC$= 0.901$), while fine-tuned BERTimbau led among supervised configurations (macro-$F_1 = 0.916$), clearly outperforming the SVM baseline (0.837). CoT regularized model behavior without improving peak performance, and RAG provided a more evidence-grounded error profile. These results confirm that inference strategy and error distribution are as consequential as aggregate metrics for interpreting model quality under distribution shift.

Beyond the specific corpus and language, three findings carry broader implications. First, parameter scale alone does not determine performance: instruction-tuned models in the 7–9B range consistently outperformed larger alternatives, suggesting that alignment quality is a stronger predictor of effectiveness than raw capacity. Second, CoT and RAG serve distinct functional roles: CoT reduces inter-model behavioral variance, while RAG shifts errors toward more interpretable failure modes. Third, the unified OOD protocol proposed here, which strictly isolates the test set from training data, retrieval indices, and prompt design, constitutes a replicable template for low-resource languages and politically sensitive domains.

The results should be interpreted in light of key limitations. The test set is small

and restricted to a specific electoral context, limiting generalization across domains and media formats. The binary veracity scheme simplifies the real-world truthfulness spectrum, and the hardest instances likely correspond to texts near the FAKE/REAL boundary; practical moderation systems would require finer-grained labels for meaningful human-in-the-loop decisions. At the methodological level, LLM outputs remain sensitive to prompting conditions — a single prompt per strategy was adopted, and prompt variation may affect results — generated justifications may be post hoc, RAG performance depends on retrieval quality, and it cannot be ruled out that the evaluated LLMs were pre-trained on data overlapping with the corpora used here, which may confer an implicit advantage in zero-shot settings. Ethically, false positives risk reputational harm and constraints on freedom of expression; the proposed framework should therefore be understood as a reproducible research instrument rather than an autonomous moderation solution.

Acknowledgments

This work was supported by CAPES (grant number 88887.671481/2022-00), Anatel (National Telecommunications Agency) and TRE-GO (Regional Electoral Court of Goiás).

Generative AI disclosure. Large language model tools, including Claude (Anthropic) and ChatGPT (OpenAI), were used for linguistic revision and support in source-code development. All scientific content, results, references, and interpretations were verified and remain the sole responsibility of the human authors.

References

Bernardi, A. J. B. (2021). *Fake news e as eleições de 2018 no Brasil: como diminuir a desinformação?* Editora Appris.

Boumber, D., Tuck, B. E., Verma, R. M., and Qachfar, F. Z. (2024). Llms for explainable few-shot deception detection. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, pages 37–47.

BRASIL (2019). Resolução tse n.º 23.610, de 18 de dezembro de 2019. Tribunal Superior Eleitoral, Brasília. Dispõe sobre propaganda eleitoral e condutas ilícitas em campanha.

Cabral, L., Monteiro, J., da Silva, J., Mattos, C., and Mourão, P. (2021). Fakewhatsapp.br: Nlp and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages. In *ICEIS*, pages 63–74.

Cantarino, F. H. S. (2024). Criação de um corpus português para auxiliar a identificação de notícias verdadeiras e falsas.

Charles, A. C., Ruback, L., and Oliveira, J. (2022). Fakepedia corpus: A flexible fake news corpus in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 37–45. Springer.

Chavarro, J. P., Carvalho, J. T., Portela, T. T., and Silva, J. C. (2023). Faketruebr: Um corpus brasileiro de notícias falsas. In *Escola Regional de Banco de Dados (ERBD)*, pages 108–117. SBC.

Cordeiro, P. and Pinheiro, V. (2019). Um corpus de notícias falsas do twitter e verificação automática de rumores em língua portuguesa. In *Proceedings of the Symposium in Information and Human Language Technology*, pages 219–228.

Couto, J. M., Pimenta, B., de Araújo, I. M., Assis, S., Reis, J. C., da Silva, A. P. C., Almeida, J. M., and Benevenuto, F. (2021). Central de fatos: Um repositório de checagens de fatos. In *Dataset Showcase Workshop (DSW)*, pages 128–137. SBC.

de Morais, J., Abonizio, H., Tavares, G., da Fonseca, 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.

Delucis, M. M., Fraga, L., Parraga, O., Mattjie, C., Ravazio, R., Barros, R. C., and Kupssinskü, L. S. (2025). Automated fact-checking in brazilian portuguese: Resources and baselines. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 137–148. SBC.

Faustini, P. and Covões, T. (2019). Fake news detection using one-class classification. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 592–597. IEEE.

Garcia, G. L., Afonso, L. C., 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. Springer.

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

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. *Anais do Encontro Nacional de Inteligência Artificial e Computacional*.

Gôlo, M., Caravanti, M., Rossi, R., Rezende, S., Nogueira, B., and Marcacini, R. (2021). Learning textual representations from multiple modalities to detect fake news through one-class learning. In *Brazilian Symposium on Multimedia and the Web*, pages 197–204.

Herculano, A., da Silva, L., Fernandes, D., and Rego, A. (2025). Uma avaliação comparativa entre o deprebertbr e modelos de linguagem para classificação de textos depressivos. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 535–548, Porto Alegre, RS, Brasil. SBC.

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., and Qi, P. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113. Number 20.

Kansaon, D., de Freitas Melo, P., Zannettou, S., and Benevenuto, F. (2025). From fake news to real protests: Whatsapp's role in brazilian political coordination. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1007–1020.

Li, Y., Jiang, B., Shu, K., and Liu, H. (2020). Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation.

Liu, B., Wang, A., and Xia, C. (2025). Interpretable chinese fake news detection with chain-of-thought and in-context learning. *IEEE Access*.

Lopez-Joya, S., Diaz-Garcia, J. A., Ruiz, M. D., and Martin-Bautista, M. J. (2025). The blueprint of a new fact-checking system: A methodology to enrich rag systems with new generated datasets. *Computers and Electrical Engineering*, 128:110746.

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 *Dataset Showcase Workshop (DSW)*, pages 138–147. SBC.

Monteiro, R. A., Santos, R., Pardo, T., de Almeida, T., Ruiz, E., and Vale, O. (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. Springer.

Moreno, J. A. and Bressan, G. (2019). Factck.br: A new dataset to study fake news. In *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web*, pages 525–527. ACM.

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. ACM.

Paiva, L., Assis, G., Amorim, A., Dias, L. G., Paes, A., and de Oliveira, D. (2025). Domínio delimitado, Ódio exposto: O uso de prompts para identificação de discurso de Ódio online com llms. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 493–506, Porto Alegre, RS, Brasil. SBC.

Patel, A., Tiwari, A., and Ahmad, S. (2022). Fake news detection using support vector machine. In *Proceedings of the 3rd International Conference on Advanced Computing and Software Engineering (ICACSE 2021)*, pages 34–38. SCITEPRESS.

Pires, V. B., Guerreiro, D., et al. (2024). Portuguese fake news classification with bert models. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 834–845. SBC.

Schreiber, A. (2022). Civil rights framework of the internet (bcrfi; marco civil da internet): Advance or setback? civil liability for damage derived from content generated by third party. In *Personality and Data Protection Rights on the Internet: Brazilian and German Approaches*, pages 241–266. Springer.

Shahi, G. K. and Nandini, D. (2020). Fakecovid – a multilingual cross-domain fact check news dataset for covid-19. In *Proceedings of ICWSM*.

Silva, F. R. M. d. (2020). Fakenewssetgen: um processo para construção de datasets que viabilizem a comparação entre métodos de detecção de fake news. Master's thesis, Instituto Militar de Engenharia (IME).

Silva, R. M., Amamou, H., Ferraz, L. B. S., da Silva, F. K. A., and Avila, A. R. (2025). Fake news detection in portuguese under large language model-generated content. *Journal of the Brazilian Computer Society*, 31(1):1149–1166.

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