Automated Fact-Checking in Brazilian Portuguese: Resources and Baselines

Marcelo M. Delucis*, Lucas Fraga*, Otávio Parraga, Christian Mattjie, Rafaela Ravazio, Rodrigo C. Barros, Lucas S. Kupssinskü

[1]MALTA - Machine Learning Theory and Applications Lab
School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul

marcelo.mussi@edu.pucrs.br

Abstract. *The spread of misinformation presents a growing societal challenge, particularly in low-resource languages such as Brazilian Portuguese (PTBR), where the scarcity of high-quality datasets limits automated fact-checking tools. In this work, we introduce translated PTBR versions of two influential English-language fact-checking datasets:* LIAR *and* AVERITEC. *These resources support multi-class veracity classification and incorporate evidence-based reasoning. We also establish baseline results for both datasets using a range of model configurations, including zero-shot and few-shot prompting with Gemma 3, and fine-tuning of encoder-based models such as mBERT, BERT-Large, and BERTimbau-Large. Across both datasets, fine-tuned encoder-based models consistently outperformed Gemma 3 in zero-shot and few-shot settings. Our results underscore the importance of task-specific fine-tuning and evidence inclusion for veracity classification in PTBR. All datasets, translation scripts, and evaluation protocols are publicly released to support further research in this area.*

1. Introduction

The rapid flow of information online has exacerbated the spread of misinformation. Misinformation creates several problems in society, such as confusion regarding scientific facts, political rumours that divide people, and a general sense of dishonesty [Domingues 2021]. Although this accelerated rate of information exchange can be beneficial [Del Vicario et al. 2016], misinformation takes the same route and misleads the public, potentially diverting police makers towards peripheral issues, undermining responses to core challenges [Southwell et al. 2019].

News-verification and fact-checking agencies attempt to curb these harms by conducting manual inspections of online information and collecting evidence to support or refute claims. Although initiatives such as these are valuable, there is an important asymmetry in the work required to generate disinformation and to debunk it [Guo et al. 2022]. Manual, labour-intensive workflows are difficult to scale up and cannot keep pace with the volume of material circulating online [Li and Chang 2023].

Automated Fact-checking powered by Natural Language Processing (NLP) has emerged as an alternative to confront misinformation spread [Guo et al. 2022]. The main advantage of employing automated fact-checking is that it is easier to scale up a tool than an entire human-centric fact-checking operation. There are many formulations for the fact-checking problem in the NLP setting; nevertheless, most of them culminate in a classification problem where we provide a statement, and the model needs to perform classification among classes that represent distinct degrees of veracity [Thorne and Vlachos 2018].

*Equal contribution.

The Automated Fact-checking task demands annotated data to train a classifier or, at least, to evaluate it. In English, such datasets include FAKE NEWS CHALLENGE [Pomerleau and Rao 2017], SEMEVAL RUMOUREVAL [Derczynski et al. 2017] and MEDIAEVAL VERIFYING MULTIMEDIA USE [Boididou et al. 2015].

Two widely used English resources exemplify complementary perspectives on automatic fact-checking. The LIAR [Wang 2017] dataset provides 12.800 short political statements from the fact agency *PolitiFact*[*], each labeled on a six-point truthfulness scale. On the other hand, AVERITEC [Schlichtkrull et al. 2023] dataset contributes with 4.568 real-world claims paired with evidence on the form of question-answer (QA) decompositions. Both works recognize that Automated Fact-Checking goes beyond the simple classification of statements and incorporates both statements and evidence in the classification pipeline.

When compared to English, data availability in low-resource languages is a limiting factor. When we search for Brazilian Portuguese (PTBR) data, we find works such as FAKEBR [Silva et al. 2020] and FAKERECOGNA [Garcia et al. 2022] datasets that, despite their importance, still treat fact-checking as a binary statement-only classification problem. More sophisticated approaches that account for more than two degrees of veracity and that utilize evidence as input, such as those found in LIAR or AVERITEC, are still lacking in PTBR.

To address this gap, our study makes two contributions: First, we create PTBR versions of both the LIAR and AVERITEC datasets; Second, we release Automated Fact-checking baseline classifiers in PTBR for both LIAR-BR and AVERITEC-BR datasets. All code and datasets are openly released in our repository[*].

Our baseline study features five distinct configurations of classifiers, spanning zero-shot and few-shot prompting of the 4 Billion parameter Gemma 3 [Kamath et al. 2025] large language model and fine-tuning of mBERT [Devlin et al. 2019], BERT-Large, and BERTimbau-Large [Souza et al. 2020] models. We achieved an F1-score of 0.44 and 0.49 for LIAR and AVERITEC, respectively, outperforming their original studies.

2. Related Work

The Fact Extraction and VERification (FEVER) dataset frames claim verification by pairing each of its 185.445 human-written assertions with sentence-level evidence from Wikipedia and assigning one of three labels: *Supported*, *Refuted* or *Not Enough Info*. Annotators also identify the minimal evidence set, yielding a Fleiss' $\kappa = 0.68$ for label reliability and multi-sentence rationales for 17% of the claims [Thorne et al. 2018]. By popularizing the pipeline of document retrieval, sentence selection and entailment classification, FEVER revealed how fragile performance remains when evidence is missing or incomplete.

In the PTBR scenario, FACTNEWS delivers the first sentence-level benchmark that annotates factuality and media bias across 6.191 sentences from 300 parallel articles [Vargas et al. 2023]. Although it enables dual-task evaluation, its labels are assigned without linking sentences to supporting sources, limiting its usefulness for evidence-aware models.

Another benchmark, the FAKE.BR pairs 3.600 fabricated news pieces with 3.600 topic-aligned legitimate news [Silva et al. 2020]. This matching exposes stylistic cues, for example, spelling-error rates, which allows Bag-of-Words (BoW) [Jurafsky and Martin 2000] models to exceed 0.97 F1 score, suggesting that the binary task can be solved with shallow lexical signals rather than reasoning. This also illustrates the lack of evidence to claims in PTBR resources.

[*] `https://www.politifact.com/`

[*] `https://github.com/Malta-Lab/automated-fact-checking-in-pt-br`

Among sources that provide metadata, the FAKERECOGNA dataset scales to 11.902 full-text items evenly split between *real* and *fake* while also providing metadata fields for each article such as title, full text, category and a class label [Garcia et al. 2022]. For real claims, the authors have scraped news from accredited news portals and Ministry of Health of Brazil[*]. Nevertheless, the FAKERECOGNA model treats the fact-checking task as a binary classification problem on an article level. Although the authors achieved a high accuracy of 94% with a Convolutional Neural Network (CNN) on FastText embeddings, they still overlook the use of evidence in their experiments.

The FAKETRUEBR balances 1.791 rumours from a fact-checking website (*Boatos.org* [*]) with 1.791 semantically similar true articles from two mainstream outlets: *G1*[*] and *Folha de São Paulo* [Chavarro et al. 2023] in a binary classification setting. This one-to-one alignment allows models to capture contemporary narratives but still lacks supporting documents beyond the paired text, and its size is modest compared with English resources.

In comparison to other PTBR datasets, FACTCK.BR stands out by compiling 1.309 human fact-checked claims annotated with a linear veracity score using the ClaimReview[*] schema [Moreno and Bressan 2019]. While the dataset records the URL of each fact-checked article, it omits the free-text rationale, leaving models without explicit evidence to learn from.

The SELFAR framework explores explainable sentence-level fack-checking by explaining zero-shot ChatGPT-4 [Achiam et al. 2023] with LIME [Ribeiro et al. 2016] and SHAP [Lundberg and Lee 2017] rationales [Vargas et al. 2024]. The authors achieve 0.85 and 0.71 F1-scores on *reliability* and *veracity* prediction, respectively, over FACTNEWS and FACTCK.BR datasets. Although the explainability of the method is promising, its effectiveness depends on external retrieval that is not grounded in the underlying datasets.

It is clear that English benchmarks still outnumber PTBR in both size and sophistication. FEVER, for instance, offers 185.000 claims, whereas current PTBR corpora, FAKERECOGNA ($\approx$ 12.000 claims), FAKE.BR ($\approx$ 7.000 claims) and FAKETRUEBR ($\approx$ 3.600 claims) remain at least an order of magnitude smaller, and sentence-level resources such as FACTNEWS ($\approx$ 6.000 sentences) and FACTCK.BR ($\approx$ 1.000 claims) are smaller still. The binary classification, absence of evidence and limited scale restrict the development of retrieval approaches or explanation-rich architectures for PTBR, underscoring the value of developing or translating multi-class, evidence-centric datasets like LIAR and AVERITEC.

Our work complements these lines by (i) releasing PTBR versions of two English datasets that combine short political claims and evidence-anchored decompositions, (ii) benchmarking fact-checking models under evidence-agnostic and evidence-aware settings.

3. Materials and Methods

Our Materials and Methods are divided into two subsections: *Datasets and Translation* where we detail the process to create LIAR-BR and AVERITEC-BR from the original datasets with Gemma 3; and *Automated Fact-Checking in PTBR* where we detail our five baselines and compare them to our rerun of LIAR and AVERITEC experiments.

[*] https://www.gov.br/saude/pt-br
[*] https://www.boatos.org/
[*] https://g1.globo.com/
[*] https://www.claimreviewproject.com/

3.1. Datasets and Translation

Both AVERITEC [Schlichtkrull et al. 2023] and LIAR [Wang 2017] datasets were translated into PTBR using the open-source Gemma 3 language model with 4 billion parameters [Kamath et al. 2025], deployed via the Ollama. The translation was performed on zero-shot prompting one field at a time with the following prompt: *"Traduza para português do Brasil. Apenas responda com a tradução, sem pensar, explicar ou comentar"*.

This prompt was used to to translate the claims, justifications, speaker information, and QA pairs (in the case of AVERITEC), while preserving timestamps, URLs, and class labels unchanged. The translated datasets were named **LIAR-BR** and **AVERITEC-BR** and are distributed in the same '.json' format as their original counterparts.

3.1.1. AVERITEC-BR

AVERITEC contains 4.568 real-world claims drawn from 50 fact-checking organizations. Each claim is decomposed into an average of 2.6 QA pairs that point to pre-claim evidence and culminate in a free-text justification; inter-annotator agreement on verdicts labels reaches a Cohen's [Cohen 1960] $\kappa = 0.619$. According to Landis–Koch scale [Landis and Koch 1977], the $\kappa \in [0.61, 0.80]$ denotes substantial reliability, that is the case for AVERITEC annotations.

The fact-checking task is modeled into a *multiclass* classification problem with four classes: *Supported*, *Refuted*, *Not Enough Evidence* and *Conflicting Evidence/Cherry-picking*; thereby capturing misleading but technically accurate statements. To provide additional information to support or refute the statement, metadata information regarding speaker, publisher, publication date and relevant locations are also available in the dataset.

AVERITEC-BR follows the AVERITEC design, where Automated Fact-Checking has two baselines: a "No Evidence" where the model has access only to the claim and to questions about the given claim; and a "Gold Evidence" where the model has access to the claim, to the questions and also to the curated answers; that constitutes the evidence of the dataset.

In its original experiment Schkichtkrull et al [2023] reported a macro F1-score of 0.17 in "No Evidence" and 0.49 in "Gold Evidence" scenarios. In our study we chose to rerun the experiments in the original data because we conjectured we could achieve better results.

3.1.2. LIAR-BR

The LIAR dataset models the Automated Fact-checking task as a *multiclass* classification problem. The distinction here is that the target variable is ordinal and has six possible valuations: *Pants on Fire*, *False*, *Barely-True*, *Half-True*, *Mostly-True*, *True*.

The LIAR-BR dataset has $12,836$ annotated political statements with metadata for *context*, *job title of the speaker*, *subject*, and *party affiliation*. Statements average 18 tokens and spans from 2007 to 2016. We are releasing a train-dev-test split with $10,269/1,284/1,283$ statements respectively. The data is almost perfectly balanced regarding the target variable.

Wang et al [2017] trained five distinct models on Automated Fact-checking classification task. The best performing model was a hybrid Convolutional Neural Network with an accuracy of 0.274, slightly above chance (0.2). We also chose to rerun all experiments in the original data because we conjectured we could achieve better results.

In order to probe whether speaker-level metadata improves Automated Fact-checking, we perform two experiments: "Statement Only", where we use only the statement as input, and "Metadata Enhanced Statement" where we use both the statement and the metadata as inputs. We report results on held-out test-set.

3.2. Automated Fact-Checking in PTBR

We adopt three strategies for Automated Fact-Checking in PTBR: *zero-shot*, *few-shot*, and *fine-tuning*: For the zero-shot and few-shot settings, we use the multilingual Gemma 3 4B model [Kamath et al. 2025]. For the fine-tuning experiments, we fine-tune three BERT variants: multilingual BERT (mBERT) [Devlin et al. 2019], BERTimbau-Large [Souza et al. 2020], and BERT-Large [Devlin et al. 2019].

In the AVERITEC-BR dataset, experiments were conducted under the two evidence availability scenarios: "No Evidence" and "Gold Evidence". In the "No Evidence" scenario, the evidence for each claim was "*No answer could be found*". As for the "Gold Evidence" scenario, each input consisted of a claim and all the pairs of the original human-curated QA. For this dataset, we report F1 scores per-class and Macro-F1 scores.

For zero-shot and few-shot evaluation using Gemma 3 we prompted the model to perform classification and provided a description of each of the available classes. In the few-shot setting, prompts were augmented with a pair of examples for each label class and formatted consistently with the targeted instance. In the fine-tuning setting, we fine-tuned the entire encoder-only transformer along with the single-layer classification head, using the claim text and contextual metadata or evidence as input.

Regarding the LIAR-BR dataset, experiments were conducted in two complementary scenario that mirror the original benchmark: "Statement Only" and "Metadata Enhanced Statement". The "Statement Only" scenario feeds only the statement to the model, and the "Metadata Enhanced Statement" scenario concatenates the statement with all its metadata.

Following the original benchmark, overall accuracy is the headline metric of LIAR-BR; we additionally report per-class F1-scores. We also add the mode of the error in order to characterize performance across the labels and to clarify how the model's mistakes are distributed. We define mode of the error (*Mode* ε) as the most frequent absolute distance between a prediction and the gold label on LIAR's ordinal six-point scale, so smaller values mean the model's typical misclassification lies closer to the true verdict. Mode error (ε) is calculated only over the misclassified instances, where $\varepsilon = 1$ indicates that the most common mistake is a one-step distance to an adjacent label (e.g., predicting *Barely True* for a *Half True* statement), while larger ε values reflect progressively more severe misclassifications.

Since LIAR's six truth labels form an ordered scale, we also computed *Cohen's* κ_w [Cohen 1968], where disagreements that are farther apart receive a heavier penalty via quadratic weights. The standard *Cohen's* κ corrects the observed agreement P_o for the agreement that could occur by chance P_e, yielding values from -1 (systematic disagreement) through 0 (chance level) to 1 (perfect consensus). The quadratic weighting preserves the usual $[-1, 1]$ range, but presents a fairer picture of agreement when, for instance, mistaking *Half True* for *Mostly True* is less severe than confusing *True* with *Pants on Fire*.

3.2.1. Zero-shot and Few-shot Prompting with Gemma 3

The Gemma 3 language model was evaluated using both zero-shot and few-shot prompting strategies, with prompt formulations specifically designed for each dataset. The scripts were constructed using Python and executed using a local instance of the Gemma 3 model via the Ollama API.

All prompts shared a common instructional prefix: *"Você é um verificador de fatos. Classifique a seguinte alegação com base apenas nas informações fornecidas. Use apenas uma das opções abaixo."* . The set of labels varied according to the dataset.

For each dataset, we evaluated two zero-shot and two few-shot configurations, totaling four prompting strategies per dataset. In AVERITEC-BR, the inputs included either (i) the claim and its associated question ("No Evidence"), or (ii) the claim, question, and corresponding answer ("Gold Evidence"). In the LIAR-BR setting, prompts were either (i) only statement ("Statement Only"), or (ii) statement and metadata ("Metadata Enhanced Statement").

For AVERITEC, the zero-shot prompting included an instruction asking the model to classify the claim based solely on its related questions. The prompt explicitly listed the four target labels and instructed the model to select the most appropriate label.

For the few-shot settings, each prompt included a balanced set of in-context examples: specifically, two manually selected instances per label, written in the same structural format as the target classification input.

These examples included the claim, a QA pair or sequence, and the expected label. They were prepended to the prompt, followed by the target claim and its questions, but without a label, requiring the model to infer the classification.

For LIAR, the zero-shot prompts followed a similar structure. When in the "Statement Only" setting, the input was solely a statement, whereas in the "Metadata Enhanced Statement" setting, both the statement and metadata were used as input. The model was then asked to classify it into one of the six LIAR's possible classes.

In the few-shot scenario, 2 labeled claims per class were inserted in the prompt. These labeled claims were presented with and without metadata ("Statement Only" and "Metadata Enhanced Statement" respectively) and served as references to guide the model's classification.

3.2.2. Fine-tuning with BERT models

For the fine-tuning, we employed mBERT, BERT-Large, and BERTimbau-Large, initialized with pre-trained weights and subsequently trained on the PTBR translated datasets. A single-layer classification head with cross-entropy loss was employed for multiclass label prediction.

All three BERT variants, the Gemma 3 zero-\few-shots trials were executed on a single NVIDIA RTX A6000. The learning rate was selected by an Optuna hyper-parameter search, varying for each BERT model for the reported runs; as the optimizer it was used the ADAMW optimiser and a REDUCELRONPLATEAU scheduler. Batch size ($64 - 512$) was adjusted per model to fit within GPU memory, and early stopping terminated training after 20 validation epochs without improvement. For AVERITEC-BR, inverse-frequency class weights compensated for label imbalance during loss and metric computation.

Our evaluation follows the original studies. On AVERITEC-BR we report per-class F1 and macro-averaged F1, whereas on the LIAR-BR we use accuracy as the primary metric; we also include per-class F1, *Cohen's* κ_w and mode errors for reference.

4. Results and Discussion

4.1. AVERITEC-BR

In the original study [Schlichtkrull et al. 2023] the authors obtained a 0.17 macro F1 in "No Evidence" scenario, and 0.49 in the "Gold Evidence" counterpart. We can see in Table 1, that our reruns on the original data outperforms in both scenarios, with mBERT achieving a 0.35 F1-score in "No Evidence" and BERT-Large 0.54 in "Gold Evidence". Our choice to perform reruns of this experiments was important because it would be unfair to compare AVERITEC-BR results to AVERITEC when our training pipeline improves upon the original work.

Table 1. Our results in the original AVERITEC dataset, with per-class F1-score test results and Macro F1-Score (upper block) and in our PTBR translation (lower block). S = Supported; R = Refuted; C = Conflicting Evidence/Cherry-picking; N = Not Enough Evidence. We highlight (bold) the highest achieving models regarding the Macro-F1 metric.

	Approach		S	R	C	N	Macro F1
AVERITEC	No Evidence	mBERT	0.43	0.71	0.13	0.12	**0.35**
		BERT-L	0.38	0.71	0.16	0.12	0.34
		Zero-Shot	0.27	0.04	0.15	0.17	0.16
		Few-Shot	0.32	0.15	0.16	0.19	0.20
	Gold Evidence	mBERT	0.49	0.73	0.21	0.58	0.50
		BERT-L	0.55	0.78	0.25	0.57	**0.54**
		Zero-Shot	0.70	0.53	0.18	0.41	0.45
		Few-Shot	0.61	0.38	0.18	0.44	0.40
AVERITEC-BR	No Evidence	mBERT	0.44	0.65	0.24	0.10	0.36
		BERTimbau	0.47	0.73	0.15	0.22	**0.39**
		BERT-L	0.39	0.68	0.08	0.14	0.32
		Zero-Shot	0.31	0.04	0.13	0.12	0.15
		Few-Shot	0.22	0.35	0.15	0.17	0.23
	Gold Evidence	mBERT	0.53	0.75	0.14	0.53	**0.49**
		BERTimbau	0.58	0.78	0.08	0.49	0.48
		BERT-L	0.40	0.76	0.17	0.46	0.45
		Zero-Shot	0.65	0.50	0.15	0.18	0.37
		Few-Shot	0.59	0.62	0.18	0.39	0.45

In the same "No Evidence" scenario, mBERT reaches 0.43 and 0.71 for the factual labels *Supported* and *Refuted*, respectively. By contrast, performance drops on the more nuanced categories: *Conflicting Evidence/Cherrypicking* with 0.13 F1-score and *Not Enough Evidence* with 0.12 F1-score. This discrepancy shows that when no external evidence is supplied, the wording of a claim alone rarely reveals cherry-picking or an absence of corroboration.

When providing models with the human-annotated QA pairs, it's evident the increase in performance: mBERT climbs to 0.50 macro F1 and BERT-Large is the best-performing model with 0.54. Gains again concentrate in the factual classes, whose F1-scores each increase, while the more nuanced labels remain with low F1-scores. In "No Evidence" scenario, Gemma 3 attains 0.16 macro F1 in the zero-shot setting, and 0.20 in the few-shot setting, but surges to 0.45 when the QA pairs are provided in the "Gold Evidence", almost matching mBERT.

In the PTBR counterpart, the benchmark exhibits the same pattern but narrows the gap between settings. Without evidence, BERTimbau-Large is able to reach 0.39 macro F1, four points above its English counterpart and five above English BERT-Large. It's possible to attribute that language-aligned pre-training helps the model exploit subtle morpho-syntactic stance markers, especially in the *Supported* class, which jumps from 0.44 to 0.58 F1. When QA pairs are provided this advantage disappears: mBERT and BERTimbau converge at 0.49 and 0.48 F1 respectively, confirming that explicit evidence, not pre-training alone, determines the capacity to correctly classify the claim. Gemma again benefits from evidence, rising from 0.15 to 0.45 macro F1. Across languages, as expected, evidence remains the decisive lever.

4.2. LIAR-BR

In its original study, Wang et al.'s [2017] sentence-level CNN achieved 0.270 accuracy without any metadata and 0.274 when all metadata were used. When we compare this to our reruns in Table 2 we see that our BERT-Large achieved nearly the same accuracy as originally reported whereas our Metadata enhanced mBERT exceeds the best previously reported scores, showing that transformer encoders can exploit contextual fields far more effectively than the earlier architecture. This corroborates our choice to do reruns of all experiments instead of relying in the accuracy metric reported by the original study.

Table 2. Our results for average accuracy in the english LIAR dataset, where we provide per-class F1-scores test results, weighted Cohen's κ and mode error (upper block), and our results in the translated dataset (lower block). We highlight the best performing models based on average accuracy (bold) and the best performing models based on weighted Cohen's κ (underscore). PF = Pants on Fire; F = False; BT = Barely True; HT = Half True; MT = Mostly True; T = True.

	Approach		PF	F	BT	HT	MT	T	Avg Acc	κ_w	Mode ε
LIAR	Statement Only	mBERT	0.18	0.23	0.28	0.22	0.24	0.29	0.25	0.23	1
		BERT-L	0.30	0.27	0.25	0.23	0.25	0.31	**0.26**	0.25	1
		Zero-Shot	0.02	0.33	0.12	0.05	0.29	0.01	0.22	0.16	1
		Few-Shot	0.18	0.06	0.09	0.12	0.22	0.07	0.14	0.12	1
	Metadata Enhanced Statement	mBERT	0.50	0.49	0.42	0.44	0.47	0.36	**0.45**	0.49	1
		BERT-L	0.60	0.44	0.38	0.41	0.47	0.38	0.43	0.46	1
		Zero-Shot	0.15	0.31	0.04	0.15	0.29	0	0.22	0.16	1
		Few-Shot	0.16	0.02	0.06	0.01	0.01	0.01	0.09	0.03	1
LIAR-BR	Statement Only	mBERT	0.21	0.25	0.21	0.29	0.23	0.23	0.24	0.23	1
		BERTimbau	0.27	0.26	0.22	0.27	0.24	0.28	**0.25**	0.25	1
		BERT-L	0.16	0.27	0.25	0.21	0.26	0.22	0.24	0.20	1
		Zero-Shot	0.10	0.18	0.25	0.30	0.03	0.01	0.21	0.11	1
		Few-Shot	0.25	0.24	0.19	0.20	0.21	0.05	0.20	0.16	1
	Metadata Enhanced Statement	mBERT	0.54	0.48	0.39	0.43	0.47	0.35	**0.44**	0.45	1
		BERTimbau	0.55	0.48	0.38	0.41	0.44	0.37	0.43	0.48	1
		BERT-L	0.37	0.44	0.39	0.42	0.40	0.29	0.39	0.37	1
		Zero-Shot	0.08	0.25	0.26	0.22	0	0	0.20	0.10	1
		Few-Shot	0.23	0.28	0.21	0.19	0.02	0.08	0.20	0.11	1

In Table 2, we can compare mBERT's performance from the "Statement Only" to the "Metadata Enhanced Statement" settings. Under this change, the largest difference appear at the extreme falsehood labels: *Pants-on-Fire* rises from an F1 of 0.18 to 0.50, and *False* climbs from 0.23 to 0.49. By contrast, the intermediate labels—*Barely True* and *Half True*—rise from 0.20 to 0.34 and 0.19 to 0.31, respectively. These improvements underscores the importance of contexts within the metadata, which helps the model identify habitual exaggerators, while finer shades of truthfulness still demand explicit evidence.

We observe the same overall trend both in PTBR and in english: the "Statement Only" is a harder multiclass classification problem when compared to the "Metadata Enhanced Statement". In the "Statement Only" setting, all three BERT backbones cluster around 0.24 accuracy, showing that only the information regarding the claim by itself is not enough to discriminate the six truthfulness labels. Once the all metadata fields are appended, in the "Metadata Enhanced Statement", accuracy rises to 0.48 for BERTimbau-Large and 0.44 for mBERT, while BERT-Large lags behind at 0.39. Weighted Cohen's κ mirrors this trajectory, climbing from about 0.23 in the "Statement Only" setting to 0.48 in the "Metadata Enhanced Statement", indicating a reduction in severe misclassifications.

We were also interested in studying how the model misclassifies labels, so we inspect

the mode of the error. Across both languages and all classification backbones the most common error is 1, meaning that when classifiers do mislabel a statement they are typically only one label away from the ground truth. This is interesting because the target variable in this classification setting is ordinal, meaning that the gap of one between predicted and actual class is less problematic than a gap of two or more.

The larger gain shown by BERTimbau-Large may be best explained by its pre-training on PTBR corpora: its vocabulary and contextual embeddings align natively with LIAR's PTBR counterpart, giving it an advantage over the multilingual (mBERT) and english models (BERT-Large), independent of any special sensitivity to the metadata fields.

A surprising finding in this experiment is that Gemma 3 underperformed when compared to other models. If we inspect accuracy alone we would be fooled into thinking that Gemma is random-guessing, however it also commited more errors by one class when compared to others. The zero-shot prompt in the "Metadata Enhanced Statement" scenario, reaches only 0.22 average accuracy and $\kappa_w = 0.16$, while the few-shot variant drops to 0.14 and 0.12. The PTBR benchmark scores track closely: the highest accuracy is 0.21 in the "Statement Only" scenario and the zero-shot setting, with κ_w never rising above 0.16 across all Gemma 3 scenarios and settings. These findings confirm that, on this task, the Gemma 3 model, with 4 billion parameters, used purely in zero- or few-shot mode cannot match the performance attainable through gradient-based fine-tuning.

4.3. Cross-dataset observations

These two benchmarks illuminate different constraints on factuality models. For LIAR in either language, the limiting factor is contextual sparsity: with only the statement, both mBERT and BERT-Large hover near 0.25 accuracy and weighted $\kappa_w \approx 0.23$. Including the full set of metadata nearly doubles accuracy: on the english set mBERT rises from 0.25 to 0.45 accuracy with κ_w increasing from 0.23 to 0.49, while on the PTBR set BERTimbau-Large climbs from 0.25 to 0.43 accuracy and κ_w from 0.25 to 0.48. At the same time the *Mode* ε falls from 1 to 0, indicating that most predictions shift from "one label off" to an exact match with the ground-truth once metadata fields are available.

AVERITEC-BR tells a complementary story: here the bottleneck is evidential, not contextual. Without the human-annotated QA pairs, macro-F1 sits in the mid 0.30s (0.35 for mBERT, 0.39 for BERTimbau-Large). Supplying the human-annotated answers raises these metrics to roughly 0.50 macro-F1, with BERT-Large peaking at 0.54 macro-F1. Yet the intermediate labels *Conflicting Evidence/Cherry-picking* and *Not Enough Evidence* remain below 0.25 F1, indicating that even perfect evidence does not fully resolve nuanced cases. PTBR pretraining helps only in the absence of evidence, where BERTimbau-Large outperforms BERT-Large in "No Evidence", but the advantage disappears once the QA pairs are included, confirming that explicit evidence, not subtle linguistic alignment, determines the upper bound.

Turning to the experiments using the Gemma 3 model, in the zero-shot setting, the model achieves just 0.22 average accuracy on english LIAR and 0.21 on our PTBR counterpart, with κ_w never rising above 0.16; in-context examples, in the few-shot setting, do not help, and metadata leaves these values essentially unchanged. Despite the low accuracy, the mode of misclassification remains $\epsilon = 1$ in every LIAR configuration, evidencing that the model still usually lands one category away from the truth.

In summary, metadata supplies the contextual cues that lifts performance on LIAR, whereas the human-annotatted QA pairs provides the decisive factual support on AVERITEC. Even so, the models continue to underperform on the subtle categories, such as *Conflict-*

ing Evidence/Cherry-picking and *Not Enough Evidence*, underscoring that additional work is needed to handle these distinctions.

It is important to highlight that both LIAR-BR and AVERITEC-BR datasets designs Automated Fact-checking as a classification task with a information bottleneck. However there is a crucial distinction, in LIAR-BR the bottleneck is *context* whereas in AVERITEC-BR it is *evidence*. This distinction appears harmless at first, but we should note that *evidence* to support or refute a claim is a superior form of information to add in a fact-checking process. To train a encoder-only model on context information only could lead to biases against political parties or venues of information. A fairness evaluation of this dataset is deferred to future work.

5. Limitations

A key limitation is the U.S.-centricity of the LIAR-BR and AVERITEC-BR corpora. Their focus on North American political discourse overlooks key Brazilian misinformation topics (e.g., local elections, agribusiness, regional health myths), underscoring the need for datasets from local sources. For our experiments, we performed translation and zero/few-shot classification using Gemma-4B. The 4B variation of Gemma was chosen due to computational constraints.

6. Conclusions

Automated Fact-checking is an important tool to confront missinformation. However, the development of Automated Fact-checking in PTBR is hindered by the lack of resources. We introduced LIAR-BR and AVERITEC-BR and provided baseline results that span zero-\few-shot prompting and supervised fine-tuning for Automated Fact-Checking in PTBR.

In the six-way LIAR-BR Fact-checking task, transformer models trained only on claim text stagnate at $\approx 25\%$ accuracy, but concatenating metadata almost doubles performance, evidencing that contextual cues are vital when no external evidence exists. In the four-way english AVERITEC-BR task, the decisive driver is factual grounding: supplying the gold QA pairs lifts macro-F1 from the mid-0.30s to 0.50 both in PTBR and english languages.

The PTBR pre-training allowed BERTimbau-Large to outperforms BERT-Large on LIAR-BR and AVERITEC-BR when evidence is absent, yet confers no additional benefit once either metadata or human annotatted QA are in place. Gemma 3 underperfomed in comparisson to fine-tuned models, showing that fine-tuning is still necessary for reliable factuality check.

The translated corpora and baselines we release close part of the resource gap for PTBR fact-checking, but the study also underscores two open challenges: (i) building native, large, balanced PTBR datasets that capture local misinformation patterns, and (ii) addressing the persistent weakness on nuanced labels, such as *Conflicting Evidence* and *Not Enough Evidence*, which remain hard even with human annotated QA pairs.

Our work highlights these challenges while laying a solid foundation for advancing automated fact-checking in PTBR and inviting the community to build on our resources and baselines. Future work should couple retrieval with explanation-centered evaluation and explore larger multilingual LLMs to push precision on these classes where it is hard to disambiguate.

7. Acknowledgments

This study was financed in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil - (CNPq) - Grant Number: 443072/2024-8, by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Finance Code 001 and by the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS, grant nr. 25/2551-0000891-3).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Boididou, C., Andreadou, K., Papadopoulos, S., Dang Nguyen, D. T., Boato, G., Riegler, M., Kompatsiaris, Y., et al. (2015). Verifying multimedia use at MediaEval 2015. In *MediaEval 2015*, volume 1436. CEUR-WS.

Chavarro, J., Carvalho, J., Portela, T., and Silva, J. (2023). FakeTrueBR: Um corpus brasileiro de notícias falsas. In *Anais da XVIII Escola Regional de Banco de Dados*, pages 108–117, Porto Alegre, RS, Brasil. SBC.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, 70(4):213–220.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559.

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., and Zubiaga, A. (2017). SemEval-2017 Task 8 RumourEval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Domingues, L. (2021). Infodemia: uma ameaça à saúde pública global durante e após a pandemia de covid-19. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, 15(1).

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*, pages 57–67. Springer International Publishing.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Jurafsky, D. and Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Practical Resources for the Mental Health Professionals Series. Prentice Hall.

Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., et al. (2025). Gemma 3 Technical Report. *CoRR*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

Li, J. and Chang, X. (2023). Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media. *Information systems frontiers*, 25(4):1479–1493.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Moreno, J. a. and Bressan, G. (2019). FACTCK.BR: A new dataset to study fake news. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, WebMedia '19, page 525–527, New York, NY, USA. Association for Computing Machinery.

Pomerleau, D. and Rao, D. (2017). Fake news challenge. Available at: `http://www.fakenewschallenge.org/`.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Schlichtkrull, M., Guo, Z., and Vlachos, A. (2023). AVERITEC: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

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

Southwell, B. G., Niederdeppe, J., Cappella, J. N., Gaysynsky, A., Kelley, D. E., Oh, A., Peterson, E. B., and Chou, W.-Y. S. (2019). Misinformation as a misunderstood challenge to public health. *American journal of preventive medicine*, 57(2):282–285.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, pages 403–417. Springer International Publishing.

Thorne, J. and Vlachos, A. (2018). Automated Fact Checking: Task formulations, methods and future directions. arXiv:1806.07687 [cs].

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

Vargas, F., Jaidka, K., Pardo, T., and Benevenuto, F. (2023). Predicting sentence-level factuality of news and bias of media outlets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206. INCOMA Ltd., Shoumen, Bulgaria.

Vargas, F., Salles, I., Alves, D., Agrawal, A., Pardo, T. A. S., and Benevenuto, F. (2024). Improving explainable fact-checking via sentence-level factual reasoning. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 192–204. Association for Computational Linguistics.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.