Retrieval-Augmented Generation with Small Language Models for Fake News Detection

Lucca Baptista Silva Ferraz[1], Jhúlia de Souza Leal[1], Anderson Raymundo Avila[2], Thiago Alexandre Salgueiro Pardo[1], Fernando Batista[3,4], Renato Moraes Silva[1],

[1] Interinstitutional Center for Computational Linguistics (NILC),
Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil

[2] Institut national de la recherche scientifique (INRS-EMT),
Université du Québec, Montréal, Québec, Canada

[3] Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal

[4] INESC-ID, Lisboa, Portugal

luccaferraz@usp.br, jhuliadsl@usp.br, anderson.avila@inrs.ca,
taspardo@icmc.usp.br, fernando.batista@iscte-iul.pt, renatoms@icmc.usp.br

Abstract

The spread of online misinformation has made fake news detection an essential tool for mitigating its negative impact, but many studies often disregard the temporal information, and existing datasets become outdated as news evolve. Some modern solutions using Retrieval-Augmented Generation (RAG) can solve the problem of unseen news events by providing context to the models. However, there are no studies evaluating the feasibility of web searches to attain context to decide whether a news article is true or not. This work aims to address this gap by conducting a comparative study between RAG-based solutions, traditional fake news classification methods, and deep learning-based methods. The results show that although RAG is a modern and promising technique, it cannot outperform techniques already adopted in the literature.

1 Introduction

Social media platforms have transformed the way we produce and consume information, with people switching from passive readers to content creators and spreaders (Jin et al., 2022). While providing easy and fast access to information, entertainment, and agile communication, these platforms have also expanded the ability to generate and spread misinformation on a massive scale (Guo et al., 2025). As shown in Vosoughi et al. (2018), fake news was 70% more likely to be shared on Twitter when compared to true news. According to the authors, that is because false news tends to be perceived as more novel than true news and people are more inclined to retweet novel information. The authors also showed that the largest sharing volumes for false news reached 1.000–100.000 users, whereas true stories rarely exceed 1.000. Hence, fake news detection can be an important asset against the spread of misinformation.

A common issue in developing fake news detection systems is the limited availability of linguistic resources for certain languages. While the vast majority of fake news datasets are developed in English, a critical gap remains for low-resource languages, such as Portuguese. Research in English benefits from several robust corpora covering diverse domains such as politics, medicine, product reviews, and social topics. Portuguese, on the other hand, still relies on a significantly smaller amount of annotated data. This resource constraints directly affect the effectiveness and development of countermeasure solutions to misinformation in the Portuguese language (Silva et al., 2024).

Nevertheless, numerous solutions have been proposed for detecting fake news, with research evolving from classical approaches (Khanam et al., 2021; Silva et al., 2020) to deep learning methods (Garcia et al., 2022; Hu et al., 2022; Mridha et al., 2021; Khan et al., 2021). Large language models (LLMs), which encode a vast amount of language and world knowledge, have shown impressive performance in several tasks, but their potential for fake news detection remains underexplored (Hu et al., 2024). One critical limitation of such models is the cost of incorporating new knowledge. For instance, as topics targeted by misinformation evolve, such models can become outdated (Mu et al., 2023). Retrieval-Augmented Generation (RAG) can alternatively mitigate this problem by providing external and

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 120–130

up-to-date context during LLMs' inference.

Another concern regarding fake news detection is the temporal aspect of news events (Guimarães et al., 2021). Classical detection studies fail by disregarding the temporal order of news articles. These studies often employ validation techniques such as holdout or k-fold cross-validation, which randomly shuffle the corpus (Silva et al., 2020; Garcia et al., 2022, 2024). This practice can inadvertently introduce future data into the training set relative to the test set, resulting in data leakage. Consequently, the reported performance results in such studies tend to be overly optimistic. As a consequence, the models tend to perform worse on new data and performance can decay through time (Kapoor and Narayanan, 2023).

To address these issues, we propose to evaluate two RAG-based solutions for Portuguese fake news detection, along with a comparative analysis of their performance against traditional machine and deep learning methods. RAG appears promising because it enables access to external and up to date information, mitigating the natural aging of datasets and providing an evaluation closer to real-world usage scenarios. We also explore Small Language Models (SLMs) within the RAG architecture as a lightweight and low-cost computational alternative to full-scale LLMs.

In this way, this study aims to understand whether smaller models, combined with RAG's contextual retrieval capability, can achieve competitive results without compromising interpretability or technological accessibility. Finally, by integrating real-time contextual information through web searches, the study seeks to investigate the impact of such dynamic updating on model performance, proposing a more adaptable and enduring paradigm for fake news detection in linguistically and informationally evolving environments.

In summary, our study is guided by the following main research questions:

- How effective and viable is a RAG-based architecture for fake news detection compared to traditional and deep learning methods?
- What is the impact of integrating real-time contextual information, obtained via web search, on the performance of RAG models?
- Can RAG models with SLMs outperform established methods in the literature, justifying a low-cost solution?

2 Related work

Fake news detection has evolved, driven by the understanding that misinformation is a dynamic, contextual, and temporal phenomenon. A landmark effort in this direction is FakeNewsNet (Shu et al., 2020), which integrates the news text content, its social dissemination context, and spatial and temporal dimensions. Despite this richness, the experiments reported in the paper use a random 80/20 split of the dataset, resulting in an essentially static reading of a changing problem and potentially overestimating performance compared to real-world scenarios, where writing patterns, topics, and sources of misinformation evolve over time.

It was precisely this gap that led part of the research community to investigate time-sensitive approaches. Zhang et al. (2023) propose a real-time architecture that treats news as a continuous stream, processing data in sliding windows, and combining event and topic extraction with topic merging to reduce dimensionality without losing semantics. By operating in this way, the method tracks the evolution of events and approaches temporally realistic evaluations, demonstrating gains in both efficiency and detection compared to baseline models.

With the rise of LLMs, a new line of research has emerged by incorporating RAG to overcome the limitations of supervised classifiers, retrieving web evidence prior to decision-making. The study by Nezafat and Samet (2024) exemplifies this advancement by coupling Mixtral-8×7B with a RAG pipeline that queries the Google Search API in real time. The authors argue that this integration reduces the tendency to generate inaccurate information by grounding the model in verifiable sources. Improvements of up to 23 percentage points are achieved compared to non-RAG LLMs. Nonetheless, the authors note failures and an increase in false positives, showing that dependence on external sources can introduce noise, which led the authors to pre-filter the retrieved results. Moreover, such architecture entails high computational costs.

Other authors have questioned the need for large-scale models for this task. Raza et al. (2025) show that, although LLMs offer advantages in generalization, smaller specialized models such as those based on BERT (Bidirectional Encoder Representations from Transformers) remain superior in accuracy and cost-effectiveness when the task is solely to determine whether a news item is true or false. The study further underscores that label quality

(e.g., AI-annotated data with human review) can be more decisive than model size, indicating that "bigger is not always better" for fake news detection.

Most of the cited works use data in English. They employ context-rich datasets, temporally aware methods, and RAG-based approaches, yet none were designed for Portuguese. Research in this language remains more limited: there are few studies with chronological evaluation and almost no protocols comparable to those used for English. This study aims to address this gap.

Taken together, these studies reveal a trajectory of advances and convergences, and it is at this intersection that the present work is situated. We evaluate, in chronological order, the use of RAG with SLM for fake news detection in Portuguese, focusing on practical feasibility, cost, and stability. The study compares pipelines with and without web retrieval and proposes a reproducible protocol that mitigates biases from static evaluations. We investigate under which conditions, and at what cost, RAG might provide advantages compared to other methods, aiming to explore a low-cost, updatable approach suitable for real-world scenarios.

2.1 A RAG-Based Approach for Fake News Detection

The RAG framework was introduced in Lewis et al. (2020) as a way to extend LLMs' knowledge without updating their internal parameters. Instead of relying solely on the information encoded in model weights, RAG retrieves relevant documents at query time and incorporates them into the generation process. As illustrated in Figure 1, the framework is based on the so-called external knowledge, and consists of two stages: a retrieval step, which selects relevant documents from the external source, and a generation step, in which the LLM produces an answer conditioned on the retrieved context. We examine two forms of RAG for fake news detection, which we explain next.

2.1.1 Closed-domain Pipeline

In the closed setting, the system retrieves context only from the training portion of the dataset stored in a vector database, as shown in Figure 1a. The training set follows the temporal holdout split described in Section 3.2. Thus, in the closed-domain RAG setting, the external knowledge used to provide context consists exclusively of chronologically prior news articles from the training split. For the closed approach, training samples with their respective labels, y, are indexed and maintained in a vector database. The collection of chunked documents is denoted Z. When a query x is received, the retriever selects the top-k documents based on similarity between the encoded query $q(x)$ and the document embeddings $d(z)$ for $z \in Z$. The generator then produces the output $p_\theta(x_i \mid [x_{1:i-1}, z_d])$ based on a contextual prompt comprising x and the retrieved documents z_d. The context comprises the chunk along with relevant metadata such as label, date, and title. When available, the document's category (e.g., sport, culture) is also included.

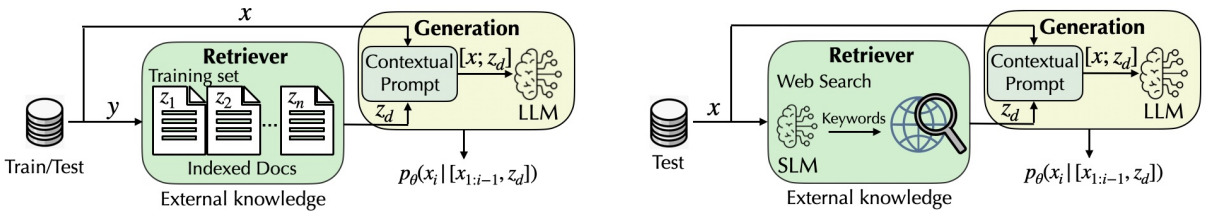
Figure 2 presents the prompt used for the closed approach.

2.1.2 Open-domain Pipeline

The open-domain RAG pipeline uses an SLM to extract keywords from each news article and formulate a search query that summarizes the event. We explored different prompting strategies, with a one-shot example proving to be the most reliable for inducing the desired query structure. The prompt shown in Figure 3 was used.

Once the queries are generated, a web search retrieves potentially related news articles to serve as external context. For each retrieved URL, a request is made and the HTML of the pages is extracted to obtain the news content along with metadata such as the title, URL, snippet, and available date. Figure 1b illustrate the entire process. Similarly to the closed-domain approach, the final stage of the open-domain pipeline performs model classification and evaluation using the contextual prompt formed by the user query x and the retrieved documents z_d.

Note that the prompt strategy shown in Figure 2 allows the LLM to use either its internal knowledge acquired during pre-training or the external information provided through the retrieved context. We also tested a second prompt strategy in which we explicitly instructed the model not to use its internal knowledge, relying solely on the supplied context for classification. Our goal was to assess whether the model's internal knowledge influences classification, since some news events may have been part of the data used during pre-training. However, both strategies produced essentially the same results, suggesting that the models did not strictly follow the instructions or that the contextual information alone was adequate for the task. Therefore, this paper reports only the results obtained with the prompt strategy illustrated in Figure 2.



(a) Closed setting.

(b) Open setting.

Figure 1: Retrieval-augmented generation with external knowledge based on (a) training data and (b) web search.

```
### SISTEMA ###
Você é um verificador de fatos. Avalie a notícia alvo. Use o CONTEXTO apenas se for claramente sobre o mesmo fato/entidade/tempo da notícia. Se o contexto não corresponder, IGNORE-O e diga isso na justificativa.

### CONTEXTO ###
{context}

### NOTÍCIA ###
{question}

### FORMATO DA RESPOSTA (TAGS) ###
<finalAnswer>
REAL ou FAKE
</finalAnswer>

<justificativa>
Em até 3 frases.
</justificativa>
```

Figure 2: Prompt (in Portuguese) for RAG approach.

```
### SISTEMA ###
Você é um agente automatizado especializado em detecção de fake news.

### TAREFA ###
Você recebe o texto de uma notícia para verificar se ela é verdadeira ou falsa. Extraia um único ponto-chave factual que seja verificável na web e elabore uma consulta na web sobre o fato. Preserve nomes próprios, números e locais.

### FORMATOS DE SAÍDA ###
Fato extraído.
<extractedText>
Consulta curta e objetiva que você elaborou.
</extractedText>

### EXEMPLO DE CONSULTA ###
Texto: pesquisa relaciona consumo de ultraprocessados ao aumento de obesidade no brasil.

Fato extraído: Há uma pesquisa que relaciona o consumo de ultraprocessados ao aumento da obesidade no Brasil.

<extractedText>
Pesquisa Brasil ultraprocessados aumento obesidade.
</extractedText>
```

Figure 3: Prompt (in Portuguese) to extract queries from the news articles.

3 Methodology

In this section, we outline the experimental setup adopted in this study.

3.1 Datasets

This study uses the FakeRecogna (Garcia et al., 2022) and Fake.Br (Monteiro et al., 2018; Silva et al., 2020) datasets for its experiments. They were chosen over other Portuguese-language datasets because many alternatives either cover only a short time span or lack publication dates. Although FakeRecogna 2.0 (Garcia et al., 2024), a recent dataset, includes publication dates, we did not use it because it does not provide the original news articles, but only their summarized versions.

We also attempted to evaluate European Portuguese fake news datasets, such as FakePT (Rodrigues, 2020) and the dataset proposed by Afonso and Rosas (2024). However, although FakePT includes date information, much of its content consists of fact-checking explanations about why the news is considered fake, rather than the news articles themselves. Meanwhile, the dataset proposed by Afonso and Rosas (2024) does not include the publication dates of the news articles.

Figure 4 presents the temporal distribution of the Fake.br and FakeRecogna datasets. Fake.br contains news from June 2009 to July 2018. The authors of FakeRecogna state that the news span from 2019 to 2023; however, the version we obtained includes articles from February 2012 to May 2023. For clarity of visualization, we restricted the time range displayed in the figure, as the number of news items outside this interval is very small. We observe that the news articles are concentrated within relatively short time spans, and there is also an imbalance between the number of fake and true news articles in certain months.

3.2 Validation and performance measures

For model evaluation, we applied a temporal hold-out validation. Specifically, the news articles were first ordered by their publication dates. The earliest 80% were used for training, while the most recent 20% were reserved for testing.

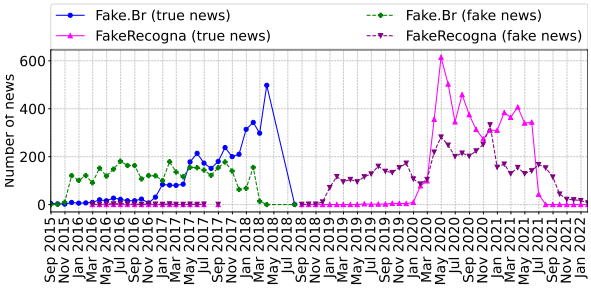


Figure 4: News published by month.

To compare model performance, we used the following well-known classification metrics: accuracy, recall, precision, and F1.

3.3 Baseline Methods

For a comprehensive evaluation of the RAG approach, we compared it against a set of classical and deep learning text classification methods that represent different learning paradigms, including a state-of-the-art contextual method (BERTimbau large) (Souza et al., 2020), a feed-forward neural network (multi-layer perceptron – MLP) (Rumelhart et al., 1986), a margin-based classifier (support vector machine – SVM) (Boser et al., 1992; Cortes and Vapnik, 1995), an ensemble of decision trees (random forest – RF) (Breiman, 2001), and a probabilistic baseline (multinomial naive Bayes – MNB) (McCallum and Nigam, 1998).

For BERTimbau, we fine-tuned the model for the fake news classification task. To optimize the hyperparameters, we used Optuna (Akiba et al., 2019) to define the learning rate {1e-6, 1e-4} and the weight decay {1e-2, 0.3}. Following the recommendations of Zhang et al. (2020), we reinitialized the top layer of the pre-trained model and opted for AdamW as the optimizer to mitigate instabilities commonly observed during fine-tuning on small datasets.

For the classical machine learning models, we also employed Optuna to optimize their key hyperparameters. Specifically, we tuned the number of trees for RF, the regularization parameter for SVM, the hidden layer size and learning rate for MLP, and the smoothing parameter for MNB. All models were implemented using the `scikit-learn` library (Pedregosa et al., 2011).

As we aim to evaluate RAG using SLMs, both because this combination remains underexplored and SLMs are more applicable than LLMs, which require significantly more infrastructure, we conducted experiments with Gemma 12B (Team, 2025), Mistral Nemo 12B (Team, 2024), and Sabiazinho 3 (Abonizio et al., 2024), an SLM trained with a focus on Brazilian Portuguese. The number of parameters of Sabiazinho 3 has not been publicly disclosed in the available technical documentation (Abonizio et al., 2024). Therefore, we report the model without specifying its parameter count.

We also intended to evaluate Qwen (14B) (Yang et al., 2025) and *DeepSeek-R1* (14B) (DeepSeek-AI, 2025). However, despite multiple prompt adjustments, these models did not consistently follow the output constraints required by our evaluation protocol (i.e., producing a single class label). This prevented us from comparing them to the other models; therefore, we do not report their results.

3.4 Text representation

In the tests of traditional methods (MLP, SVM, RF, and MNB), we adopted a bag-of-words (BoW) representation and applied a TF-IDF (term frequency-inverse document frequency) transformation fitted on the training data to weight the tokens in each document. Although the MNB model was originally defined for integer term counts, Rennie et al. (2003) showed that it can be used with fractional term weights, such as TF-IDF.

In the experiments with BERTimbau large, we used its internal contextual representations. The experiments with RAG also used BERTimbau large.

3.5 RAG Design

In this work, we used LangChain[1] to implement the RAG solution described in Section 2.1, a library that provides a modular interface for integrating LLMs with retrieval mechanisms. Our system utilizes the FAISS (Douze et al., 2025) vector database to store and retrieve external knowledge. We used its default similarity measure. To improve retrieval quality, the documents were split into chunks of 400 tokens with a 20% overlap. We tested different configurations, and empirically chose these values.

For both open- and closed-domain RAG pipelines, we selected the top-5 chunks in the retrieval step. The embedding model used for document indexing and retrieval was BERTimbau large.

For query generation in the open-domain RAG pipeline (Section 2.1.2), the SLM we used was Gemma 3, with 12 billion parameters. The use of

[1]LangChain. Available at `https://www.langchain.com/`. Accessed on: March 16, 2026.

Qwen 3 14B and Llama 3 8B was also evaluated, but both models consistently hallucinated, either deviating from the desired format or producing outputs inconsistent with the task description. We used the same SLM for query generation in all experiments to ensure that all downstream online models receive the same external knowledge for factual verification, maintaining fairness and consistency in the evaluation pipeline. Figure 5 illustrates an example of a query extracted from a FakeRecogna news article.

In addition, we conducted an empirical qualitative inspection of the queries generated by the SLM in order to assess whether the extracted keywords were meaningful for information retrieval. This inspection consisted of manually reviewing a substantial subset of the generated queries and comparing them with the content of the corresponding news articles, verifying whether the queries preserved key elements such as person names, locations, institutions, and key events described in the articles.

O Ministério da Economia confirmou nesta segunda-feira, 1º de março, a saída do secretário de Coordenação e Governança das Empresas Estatais (SEST), Amaro Gomes, que será substituído pelo secretário-adjunto, Ricardo Faria. Em nota, a pasta disse que a saída de Gomes foi "decisão de cunho pessoal" para assumir "novos desafios no setor privado". Mais cedo, Gomes negou ao Broadcast (sistema de notícias em tempo real do Grupo Estado) e ao jornal O Estado de S. Paulo que tenha pedido para deixar o cargo por causa de insatisfação com a política do governo Jair Bolsonaro depois do anúncio da troca de comando da Petrobras. Em entrevista, ele disse que foi uma "infeliz coincidência", que acertou a sua saída do cargo com o secretário Especial de Desestatização, Diogo Mac Cord, em janeiro e que de lá para cá trabalha na sua transição.

(a) Original news article (in Portuguese).

Amaro Gomes substituído Ricardo Faria Secretaria de Coordenação

(b) Query extracted from the news article (in Portuguese).

Figure 5: Example of query extraction from a FakeRecogna news article.

For the web search step in the open-domain RAG pipeline, we used the Google Custom Search API[2] due to its free usage tier and its ability to filter Google pages by publication date. For each query, the ten most relevant results are requested. A large number of websites are searched, considering the scraping impediments imposed by the pages (e.g., crawler blocking, paywalls, CAPTCHAs, rate limits). For each news item, the title, URL, snippet, and date available on Google News are extracted. The website scraping step is done using the Trafilatura (Barbaresi, 2021) and Beautiful Soup[3] libraries.

To illustrate the complexity of the data collection process, an analysis of the FakeRecogna dataset showed that, out of 22,201 pages targeted for scraping, 4,019 resulted in access or parsing failures. Despite the errors, only two test news articles were retrieved without context for FakeRecogna, while this did not occur for Fake.Br. The number of web searches (10 sites) was sufficient to guarantee this.

	RAG Web		RAG Local	
Dataset	**Real**	**Fake**	**Real**	**Fake**
FakeRecogna	0.8742	0.7769	0.8323	0.7589
FakeBR	0.8932	0.8576	0.8432	0.8126

Table 1: Average cosine similarity between the retrieved documents and the original news articles.

To further analyze the relevance of the retrieved evidence, Table 1 presents the average cosine similarity between the retrieved documents and the original news articles for each dataset and class. The results indicate that, in our experiments, the retrieved documents exhibit higher similarity with real news articles than with fake ones.

4 Results

Tables 2 and 3 present the results for each model on the FakeRecogna and Fake.Br datasets, respectively. The scores are displayed using a grayscale heatmap, where darker cells indicate better performance. Additionally, the best results in each column are highlighted in bold. The models corresponding to the closed-domain RAG pipeline (Section 2.1.1) are referred to as *RAG_Local*, while those associated with the open-domain RAG pipeline (Section 2.1.2) are denoted *RAG_Open*. Models without either label perform classification without relying on external knowledge.

The results in Table 2 show that traditional machine-learning approaches outperform RAG-based models. Classic methods, as well as the BERTimbau large model, consistently achieve

[2]Google Custom Search API. Available at: https://developers.google.com/custom-search/v1/overview. Acessed on: March 16, 2026.

[3]Beautiful Soup. Available at: https://beautiful-soup-4.readthedocs.io/. Acessed on: March 16, 2026.

Model	Acc.	Precision	Recall	F1
BERTimbau large	**0.989**	**0.994**	0.984	**0.989**
MLP	0.933	0.949	0.915	0.932
SVM	0.929	0.955	0.903	0.928
RF	0.903	0.939	0.864	0.900
MNB	0.882	0.937	0.822	0.876
Sabiazinho3	0.842	0.826	0.870	0.848
Sabiazinho3 (RAG_Local)	0.882	0.953	0.813	0.877
Sabiazinho3 (RAG_Open)	0.867	0.944	0.789	0.860
Gemma3	0.821	0.954	0.679	0.794
Gemma3 (RAG_Local)	0.882	0.879	0.889	0.884
Gemma3 (RAG_Open)	0.849	0.896	0.793	0.841
Mistral_nemo	0.727	0.654	0.966	0.780
Mistral_nemo (RAG_Open)	0.528	0.517	**0.999**	0.681
Mistral_nemo (RAG_Local)	0.786	0.709	0.978	0.822

Table 2: Classification results for FakeRecogna dataset.

Model	Acc.	Precision	Recall	F1
BERTimbau large	**0.993**	**0.951**	**0.994**	**0.972**
MLP	0.963	0.780	0.983	0.870
SVM	0.959	0.752	**0.994**	0.854
RF	0.900	0.553	0.989	0.745
NB	0.926	0.651	0.864	0.742
Sabiazinho3	0.913	0.745	0.446	0.558
Sabiazinho3 (RAG_Local)	0.935	0.944	0.925	0.935
Sabiazinho3 (RAG_Open)	0.919	0.842	0.423	0.563
Gemma3	0.862	0.533	0.384	0.446
Gemma3 (RAG_Local)	0.930	0.909	0.627	0.742
Gemma3 (RAG_Open)	0.871	0.477	0.474	0.475
Mistral_nemo	0.859	0.522	0.380	0.440
Mistral_nemo (RAG_Open)	0.170	0.127	0.983	0.126
Mistral_nemo (RAG_Local)	0.913	0.771	0.418	0.542

Table 3: Classification results for Fake.Br dataset.

higher accuracy, precision, recall, and F1-scores than the RAG configurations. In particular, BERTimbau large exhibits near-perfect performance (Acc. = 0.989 and F1 = 0.989), followed by MLP (F1 = 0.932) and SVM (F1 = 0.928). The best RAG-based result was obtained using Gemma in the closed-domain setting (F1 = 0.884).

For the Fake.Br dataset, the analysis of the results is less straightforward. As the temporal split leads to class imbalance, the holdout period contains more true than fake news, which distorts the interpretation of some metrics. Despite this limitation, the overall behavior of the RAG variants was similar to what was observed in FakeRecogna. Notably, Sabiazinho achieved accuracy comparable to that of the traditional baselines and demonstrated more stable behavior across the other metrics, which were affected by class imbalance. Gemma in the closed-domain setting also exhibited consistent performance; however, both models were still outperformed by BERTimbau large model, which retains the highest scores.

These results suggest that lexical and distributional features provide better signals to discriminate between true and fake news, outperforming architectures that rely on RAG. This pattern allows an important conclusion: the addition of contextual information did not translate into improved classification accuracy. Instead, the RAG pipeline often introduced noise or misleading signals due to limitations at the retrieval stage, particularly when the retrieved documents were not factually aligned with the original claim.

For instance, consider a fake news claim referencing a specific Brazilian law. The RAG pipeline may retrieve documents discussing other legal cases or related legislation, but not the exact law mentioned in the claim. As a result, even though the retrieved context is thematically related, it is not factually aligned, limiting the SLM's ability to leverage it for verification. Similar challenges have been reported in prior work showing that LLMs still struggle to deal with noisy or misleading retrieved evidence in RAG pipelines (Chen et al., 2024). This misalignment was more prominent within the open-domain settings, which provided lower performance compared to the closed-domain one, except for the Mistral-based LLM.

Across the models evaluated, performance varied substantially both in terms of classification quality and in the ability to follow the prescribed response format. Among the open-source models tested, Gemma 3 exhibited the most consistent behavior. It reliably respected the output structure, showed stable decision patterns across experimental conditions, and achieved the strongest overall classification results. When including non-open-source models, however, Sabiazinho performed best overall, proving highly consistent and efficient for RAG-based pipelines. This consistency makes them particularly suitable for fact-checking workflows that rely on RAG-style conditioning and require deterministic, schema-conforming outputs.

The behavior of Mistral-Nemo differed significantly between experimental domains. In the open-domain RAG setting, its performance was notably poor. Several prompt variants were tested in an attempt to improve robustness; however, the model displayed a persistent failure mode: whenever the retrieved context did not describe the same real-world event as the target news article, it systematically classified the claim as false, even when explicitly instructed not to rely solely on context overlap

as a basis for rejection. This is particularly problematic, since fake news is often poorly documented or refers to events that never occurred; consequently, retrieval frequently returns unrelated but legitimate context. Under such conditions, Mistral-Nemo exhibits overconfidence in negative predictions.

In the closed-domain RAG configuration, the performance of Mistral-Nemo improved. When the retrieved context was semantically similar, the predictions became more balanced, approaching the performance of the baseline non-RAG models. This suggests that Mistral-Nemo benefits from inductive context but struggles when required to perform genuine factual verification. Besides the improvement, its classification consistency remains inferior to Gemma 3 and Sabiazinho 3.

In the experiments on FakeRecogna, Gemma 3 and Sabiazinho 3 models in the open-domain setting obtained an F1-score of 0.841 and 0.860, respectively. The number of false positives and false negatives (110 and 248 for Gemma; 57 and 259 for Sabiazinho) indicates a markedly asymmetric behavior: although the number of false positives is low, there is a substantial volume of false negatives. In other words, the model demonstrates a strong ability to avoid labeling true news as false (high precision) but struggles to correctly identify fake content. This pattern also happens for Fake.Br: the Sabiazinho model under the open-domain setting showed 14 false negatives and 102 false positives.

The experiments with Gemma3 and Sabiazinho 3 in the closed-domain scenario revealed superior performance across all the metrics, while the distribution of errors was more balanced, especially regarding false negatives, as evidenced by the substantially improved recall when compared to the open domain configuration. This occurs because, in this setting, the retrieved instances are not used primarily for factual verification but rather serve as labeled examples that are linguistic or semantically similar to the input text. The task then becomes closer to a conventional supervised classification, in which inference relies on semantic patterns and statistical associations between texts and labels. In practice, the retrieved context often contained news items with similar narratives that were already classified as either fake or true, providing an additional inductive signal that reinforced the model's decision-making. Thus, even in the absence of specific factual evidence about the target event, the model benefits from contextual cues derived from previously labeled samples, enabling it to identify fake news by analogy rather than direct fact-checking. This mechanism helps mitigate false negatives and leads to higher F1-scores.

In the LLM-only pipeline (without RAG), the model relies exclusively on its parametric knowledge and on the textual information contained in the news article, without accessing any external sources. As a result, its predictions are informed solely by the linguistic patterns and background knowledge in pre-training, rather than by evidence grounded in factual verification. This limitation is reflected in its performance: the LLM alone achieves results slightly worse than those of the RAG-based models, indicating that, even with imperfect retrieval, the additional contextual signals still provide useful guidance for the classification.

Overall, the results confirm that, under the evaluated setups, RAG-based approaches do not yet achieve performance competitive with traditional methods. Nevertheless, the ability to ground predictions in retrieved evidence offers an advantage: model decisions become more transparent, enabling users to inspect and contextualize classification outcomes, as shown in Figure 6. This traceability is valuable in fact-checking applications, where understanding why a decision was made is often as important as the prediction itself. However, these benefits come at a high cost given that SLM-based pipelines are computationally more expensive and less efficient than lightweight traditional models, raising practical concerns about scalability and deployment in real-world settings.

5 Final remarks

This study provides an evaluation of RAG architectures for fake news detection in Portuguese. One of the research questions evaluated was the effectiveness and feasibility of RAG-based architectures for fake news classification compared to traditional and deep learning approaches.

Our findings indicate that, despite the appeal to adopt RAG methods for incorporating up-to-date factual information, they did not outperform established baselines, especially in the open-domain setting. In our experiments, they were outperformed by established baselines such as BERTimbau or classical machine learning techniques. There is evidence that the linguistic aspects of the evaluated dataset were better explored by traditional and deep learning methods than by the combination of linguistic and factual aspects of RAG and SLM. We

Notícia de teste
circula em redes sociais uma mensagem que diz que o critério de envio das doses de vacina contra a covid-19 nas cidades do estado de são paulo foi alterado. a mensagem afirma que, a partir de agora, as doses são enviadas de acordo com o número de pessoas, de cada faixa etária, que fizeram o pré-cadastro no site vacina já, do estado de são paulo. ainda segundo a publicação, quanto maior o número de pessoas cadastradas, maior a quantidade de doses enviadas para os municípios.

Classificação
<finalAnswer>
FAKE
</finalAnswer>

Justificativa
<justificativa>
A notícia falsa sugere que a distribuição de vacinas em São Paulo é baseada no número de pré-cadastros, enquanto o contexto mostra que a distribuição é baseada no número de pessoas imunizadas contra a gripe em 2020, conforme o Programa Nacional de Imunizações. Além disso, a Secretaria Estadual de Saúde de São Paulo prioriza profissionais da linha de frente no combate à Covid-19. A vacinação no Hospital das Clínicas da FMUSP inclui médicos, enfermeiros e fisioterapeutas que têm contato com pacientes Covid, refutando a ideia de vacinação baseada em pré-cadastros.
</justificativa>

Figure 6: Example of classification and justification performed by Sabiazinho 3 on the open-domain setting (in Portuguese).

do not have enough data to rule out the use of factual data by the other models, but the fact that RAG brings new information to the models is sufficient evidence to infer that SLM receives factual updates while the others do not.

The main challenges encountered stem from the retrieval of contextually-related but not factually-aligned documents, especially in open-domain settings, which adds noise compromising classification performance. This directly addresses another research question of this study regarding the impact of integrating real-time contextual information via web search: the closed-domain RAG pipeline consistently outperformed its open-domain counterpart, likely due to its ability to exploit dataset-specific linguistic regularities.

Finally, addressing our third research question, although SLMs offer a lower-cost and more accessible alternative to full-scale LLMs, the results indicate that they are not reliable for fake news detection in this setting. Even with the additional contextual information provided by the RAG pipeline, their performance remained inferior to that of classic machine learning methods, which are even less computationally expensive. While RAG has advantages, such as enabling explainability and incorporating more recent information, substantial improvements in data quality and retrieval accuracy are necessary.

Future research may address current limitations by creating more temporally balanced fake news datasets in Portuguese, enabling clearer conclusions about the benefits of incorporating updated information into machine learning methods. The temporal dimension proved crucial in this work: considering time led to lower performance compared to classical studies such as Silva et al. (2020) and Garcia et al. (2022), underscoring the sensitivity of results to temporal distribution. It is also important to develop datasets with a better balance between factual and linguistic features. Relying too much on linguistic cues makes detection easier to bypass, as fake news can be created to imitate the style of true news. Factual content should therefore play a more important role than writing style.

One limitation of this work is that intrinsic retrieval metrics were not computed due to the dynamic and unlabeled nature of the evidence retrieved from the web. Future work may address this limitation by constructing a curated gold standard enabling isolated evaluation of the retrieval component. In addition, an oracle experiment providing the model with evidence previously validated by human experts could be used. Such an approach would help isolate potential bottlenecks, such as determining whether errors arise from the retrieval stage or from limitations of the language model.

Acknowledgments

We gratefully acknowledge the support provided by São Paulo Research Foundation (FAPESP; grants #2024/17834-6 and #2025/13608-4) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – grant #444933/2024-7). Some researchers involved in this study are also affiliated with the Center for Artificial Intelligence of the University of São Paulo (C4AI – http://c4ai.inova.usp.br/), with support by FAPESP (grant #2019/07665-4), the IBM Corporation, and the Ministry of Science, Technology and Innovation, with resources from Law No. 8,248 of October 23, 1991, under the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2024. Sabiá-3 technical report. *Preprint*, arXiv:2410.12049.

Ricardo Afonso and João Rosas. 2024. Development of a smartphone application and chrome extension to detect fake news in english and european portuguese. *IEEE Latin America Transactions*, 22(4):294–303.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory (COLT'92)*, pages 144–152, Pittsburgh, PA, USA. ACM.

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

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Corinna Cortes and Vladimir N. Vapnik. 1995. Support-vector networks. 20(3):273–297.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The Faiss library. *Preprint*, arXiv:2401.08281.

Gabriel L Garcia, Luis CS Afonso, and João P Papa. 2022. FakeRecogna: a new brazilian corpus for fake news detection. In *International Conference on Computational Processing of the Portuguese Language*, pages 57–67. Springer.

Gabriel Lino Garcia, Pedro Henrique Paiola, Danilo Samuel Jodas, Luis Afonso Sugi, and João Paulo Papa. 2024. Text summarization and temporal learning models applied to Portuguese fake news detection in a novel Brazilian corpus dataset. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 86–96, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Nuno Guimarães, Álvaro Figueira, and Luís Torgo. 2021. Can fake news detection models maintain the performance through time? a longitudinal evaluation of twitter publications. *Mathematics*, 9(22):2988.

Hao Guo, Zihan Ma, Zhi Zeng, Minnan Luo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2025. Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 228–236.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 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.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI open*, 3:133–155.

Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 5746–5754.

Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804.

Junaed Younus Khan, Md Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. 2021. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:100032.

Zeba Khanam, BN Alwasel, H Sirafi, and Mamoon Rashid. 2021. Fake news detection using machine learning approaches. In *IOP conference series: materials science and engineering*, volume 1099, page 012040. IOP Publishing.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive Bayes text classification. In *Proceedings of the 15th AAAI Workshop on Learning for Text Categorization (AAAI'98)*, pages

41–48, Madison, Wisconsin. AAAI Press/The MIT Press.

Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, and Oto A. Vale. 2018. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *13th International Conference on Computational Processing of the Portuguese Language (PROPOR'2018)*, pages 324–334, Canela, Rio Grande do Sul, Brazil. Springer International Publishing.

Muhammad Firoz Mridha, Ashfia Jannat Keya, Md Abdul Hamid, Muhammad Mostafa Monowar, and Md Saifur Rahman. 2021. A comprehensive review on fake news detection with deep learning. *IEEE access*, 9:156151–156170.

Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023. It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 736–743, Dubrovnik, Croatia. Association for Computational Linguistics.

Mohammad Vatani Nezafat and Saeed Samet. 2024. Fake news detection with retrieval augmented generative artificial intelligence. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 160–167. IEEE.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Shaina Raza, Drai Paulen-Patterson, and Chen Ding. 2025. Fake news detection: comparative evaluation of BERT-like models and large language models with generative ai-annotated data. *Knowledge and Information Systems*, 67(4):3267–3292.

Jason D. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. 2003. Tackling the poor assumptions of naive Bayes text classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*, volume 3, pages 616–623, Washington, DC, USA. AAAI Press.

João Filipe Carriço Rodrigues. 2020. Fake news classification in european portuguese language. Master's thesis, ISCTE-Instituto Universitario de Lisboa (Portugal).

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Renato Moraes Silva, Roney Lira de Sales Santos, and Thiago Alexandre Salgueiro Pardo. 2024. Detecção automática de notícias falsas. In Helena de Medeiros Caseli and Maria das Graças Volpe Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edition, book chapter 27, pages 635–653. BPLN.

Renato Moraes Silva, Roney Lira de Sales Santos, Thiago Alexandre Salgueiro Pardo, and Tiago A. Almeida. 2020. Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:1–48.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*, pages 403–417, Cham. Springer International Publishing.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Mistral AI Team. 2024. Mistral nemo. `https://mistral.ai/news/mistral-nemo`. Accessed: March 16, 2026.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chaowei Zhang, Ashish Gupta, Xiao Qin, and Yi Zhou. 2023. A computational approach for real-time detection of fake news. *Expert Systems with Applications*, 221:119656.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2020. Revisiting few-sample BERT fine-tuning. *CoRR*, abs/2006.05987.