# Desempenho alto, generalização frágil: lacunas de generalização e aprendizado por atalho em detecção de fake news em português brasileiro

*High In-Distribution Performance, Fragile Generalization: Generalization Gaps and Shortcut Learning in Brazilian Portuguese Fake News Detection*

**Guilherme O. Silva**[1], **Reinaldo A. C. Bianchi**[1]

[1]Centro Universitário FEI
São Bernardo do Campo – SP – Brasil

guilherme.oliveirasilva@gmail.com, rbianchi@fei.edu.br

***Abstract.*** *This work evaluates the generalization ability of text-classification models for fake news detection in Brazilian Portuguese, assessing seventeen configurations, from classical baselines to Transformers, on FakeRecogna 2.0. In-distribution, the best models reach macro-F1 around 0.96, with no statistically significant difference among the top architectures; under cross-dataset evaluation (Fake.br), macro-F1 drops by 27–28 points, falling to 0.68–0.70 in both transfer directions. Explainability and stress analyses reveal corpus regularities consistent with shortcut learning, treated as a hypothesis rather than a proven mechanism; strikingly, a trivial classifier using only metadata (category and year) matches the neural ensemble in-distribution.*

***Resumo.*** *Avalia-se a capacidade de generalização de modelos de classificação textual na detecção de* fake news *em português brasileiro, com dezessete configurações, de baselines clássicos a Transformers, avaliadas no FakeRecogna 2.0. Dentro da distribuição, os melhores modelos atingem F1-macro próximo de 0,96 e são indistinguíveis entre si; na avaliação cross-dataset (base Fake.br), o F1-macro cai 27–28 pontos, ficando entre 0,68 e 0,70 nas duas direções de transferência. Análises de explicabilidade e de estresse apontam regularidades do corpus compatíveis com aprendizado por atalho, tratado como hipótese; um classificador trivial que usa apenas metadados (categoria e ano) iguala o ensemble neural dentro da distribuição.*

## 1. Introdução

A circulação de *fake news* tornou-se preocupação central para a integridade informacional, com episódios marcantes em eleições e na infodemia da COVID-19 [Vosoughi et al. 2018, Soares et al. 2021]. Neste trabalho, o termo *fake news* é empregado em sentido operacional (conteúdos rotulados como falsos por conjuntos de dados específicos), não como detecção de verdade em sentido amplo [Wardle and Derakhshan 2017], e a tarefa é tratada como classificação textual supervisionada.

Vários dos trabalhos brasileiros revisados avaliam modelos por particionamento aleatório de um *único* corpus, relatando métricas agregadas sem verificação sistemática de generalização cross-dataset, limitação também reconhecida de forma mais ampla na literatura [Oshikawa et al. 2020]. Métricas elevadas podem indicar aprendizado de padrões relevantes, mas também podem refletir *artefatos* de coleta, rotulagem e pré-processamento de cada corpus. Essa possibilidade conecta-se ao *aprendizado por atalho* (*shortcut learning*): modelos internalizam correlações superficiais que sustentam o desempenho dentro da distribuição, mas o degradam fora dela [Geirhos et al. 2020].

O objetivo deste trabalho é avaliar a capacidade de generalização de modelos de classificação textual na detecção de *fake news* em português brasileiro, distinguindo o desempenho efetivo do aprendizado por atalho. Para isso, vai-se além de relatar métricas agregadas em um único corpus, combinando um protocolo estatístico rigoroso com uma bateria de diagnósticos de generalização (Figura 1).

Formula-se a seguinte pergunta de pesquisa: *em que medida um conjunto comparativo de modelos (embeddings contextuais, ensembles de CNN/LSTM/ConvLSTM e modelos de linguagem pré-treinados, PLMs) apresenta desempenho estatisticamente sustentado e generalização mensurável para detecção de fake news em português brasileiro, e quais são seus limites sob protocolos mais exigentes do que a validação aleatória em um único corpus?*

As contribuições deste trabalho são:

- *experimental*: comparação de dezessete configurações no FakeRecogna 2.0 sob protocolo estatístico unificado (bootstrap, McNemar com Holm, calibração);
- *diagnóstica*: triangulação de análise lexical, curva de aprendizado, testes de estresse, sondas de atalho, avaliação cross-dataset e explicabilidade (XAI) para examinar a natureza do sinal aprendido;
- *metodológica*: um protocolo crítico de avaliação para a tarefa (Figura 1).

O termo "atalho" é tratado como hipótese diagnóstica triangulada, nunca como mecanismo causal comprovado.

## 2. Trabalhos Relacionados

O Fake.br-Corpus, um dos primeiros corpora de *fake news* em português, reúne 7.200 notícias balanceadas (3.600 por classe), com pareamento por similaridade [Monteiro et al. 2018]. Trabalhos posteriores ampliaram a disponibilidade de dados para a tarefa, culminando no FakeRecogna 2.0, com mais de 52 mil notícias e variantes de sumarização abstrativa e extrativa [Garcia et al. 2024]; esse é o corpus central deste estudo.

Quanto aos modelos, abordagens clássicas com atributos linguísticos já alcançam alta separabilidade em corpora brasileiros de *fake news* [Silva et al. 2020, Abonizio et al. 2020], assim como redes de atenção hierárquica, embora com vantagem desprezível em texto truncado [Okano et al. 2020]. Transformers pré-treinados elevaram o desempenho intra-corpus na detecção de *fake news* em português [Fischer et al. 2022], com o BERTimbau como modelo de referência para a língua [Souza et al. 2020]. Trabalhos recentes estendem a tarefa a conteúdo gerado por modelos de linguagem [Silva et al. 2025] e aplicam técnicas de XAI a corpora brasileiros da tarefa [Vicentini et al. 2026].

A lacuna de generalização decorre do *dataset shift*, a divergência entre as distribuições de treino e teste [Quiñonero-Candela et al. 2009]: sob esse regime, a incerteza preditiva se degrada fora da distribuição [Ovadia et al. 2019], e correlações espúrias e artefatos de anotação produzem desempenho não generalizável [Geirhos et al. 2020, Gururangan et al. 2018]; o problema é agravado por vazamento de dados [Kapoor and Narayanan 2023]. Em português, há evidência direta de queda acentuada na generalização cruzada de Transformers entre corpora de *fake news* [Guedes et al. 2026] e de *concept drift* temporal [Wanderley et al. 2025]. Diferentemente de [Guedes et al. 2026], que mede a degradação cross-dataset de dois Transformers (BERTimbau e mBERT) em quatro corpora, avaliam-se aqui dezessete configurações (de baselines clássicos a PLMs) sob protocolo estatístico unificado e triangulam-se diagnósticos (lexical, curva de aprendizado, estresse, sondas de atalho e XAI) para examinar a *natureza* do sinal aprendido.

## 3. Materiais e Métodos

A Figura 1 sintetiza o protocolo completo, das etapas de dados ao veredito diagnóstico; as subseções a seguir detalham cada bloco.

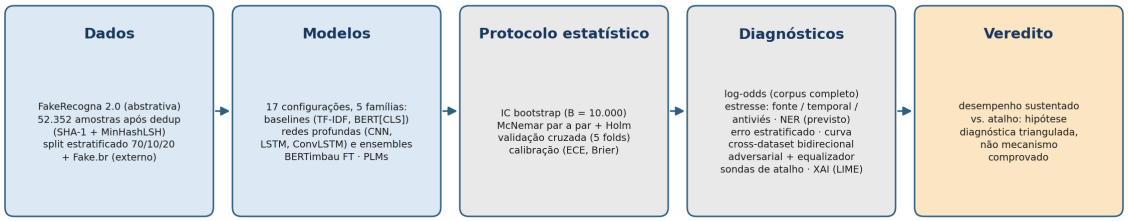



**Figura 1. Visão geral do protocolo crítico de avaliação: dados e particionamento, dezessete configurações em cinco famílias, protocolo estatístico e bateria de diagnósticos cuja convergência sustenta (ou refuta) a hipótese de aprendizado por atalho.**

### 3.1. Materiais

Os experimentos usam a variante *abstrativa* do FakeRecogna 2.0 [Garcia et al. 2024]. Após deduplicação (hash SHA-1 e MinHashLSH com limiar de Jaccard 0,85), restaram 52.352 amostras, particionadas de forma estratificada em 36.645/5.236/10.471 (treino/validação/teste; teste praticamente balanceado, 5.245/5.226). Na classe falsa, o texto é mantido sem sumarização; apenas a classe verdadeira é sumarizada, uma assimetria de construção tratada como possível atalho. Como corpus externo para avaliação cross-dataset, usa-se o Fake.br-Corpus na variante `size_normalized` (7.200 amostras) [Monteiro et al. 2018]. A Tabela 1 ilustra a tarefa com um trecho de cada classe.

### 3.2. Modelos e protocolo estatístico

As dezessete configurações comparadas abrangem cinco famílias: (a) seis baselines clássicos: TF-IDF (unigramas+bigramas) e embedding [CLS] do BERTimbau, cada um com Regressão Logística, SVM linear e MLP; (b) três redes profundas sobre embeddings por token do BERTimbau (congelado), a saber, TextCNN (doravante CNN), BiLSTM (doravante LSTM) e ConvLSTM; (c) quatro ensembles por média simples

**Tabela 1. Trechos iniciais de um exemplo de cada classe (texto como no corpus, em caixa baixa). O exemplo falso exibe o metatexto típico de checagem ("circula nas redes sociais"; cf. Tabela 5).**

| Classe | Trecho |
|---|---|
| Verdadeira | "a ssp (secretaria de segurança pública) do estado de goiás confirmou ao uol que o policial militar responsável por prender um dirigente do pt (partido dos trabalhadores) [...]" |
| Falsa | "circula nas redes sociais uma mensagem que diz que 'lula foi autorizado a dar entrevista dia 27, sábado [...]' " |

(Ens2, Ens3) e ponderada (WEns2, WEns3), com pesos otimizados na validação; (d) o BERTimbau-base com ajuste fino parcial (cabeça e últimos trinta tensores do codificador) [Souza et al. 2020]; e (e) três PLMs maiores: BERTimbau-large, XLM-RoBERTa-base e mDeBERTa-v3-base [Conneau et al. 2020, He et al. 2021].

O treinamento usa `MAX_LEN` = 200 tokens e otimizador AdamW. As redes profundas usam taxa de aprendizado $1 \times 10^{-3}$ e lote 32; o BERTimbau-base ajustado usa $2 \times 10^{-5}$ e lote 16. As redes profundas são treinadas com as sementes $\{42, 7, 2024\}$ e os modelos Transformer ajustados com $\{42, 7\}$. *Convenção de relato:* para reduzir variância, as linhas CNN, LSTM e ConvLSTM (Tabela 2) apresentam a média das probabilidades preditivas entre as três sementes (um *ensemble de sementes* de uma mesma arquitetura, distinto de Ens2/Ens3, que combinam arquiteturas diferentes); as linhas dos PLMs e do BERTimbau ajustado usam a semente principal (42) e os baselines clássicos são determinísticos. A convenção não distorce a comparação: o desvio-padrão entre sementes dos modelos profundos é $\leq$ 0,0024 em F1-macro e, nos Transformers, a média entre sementes difere do valor de semente única em $\leq$ 0,0028.

O protocolo estatístico tem como métrica principal o F1-macro. A escolha justifica-se pelo teste quase balanceado, pelo tratamento simétrico das classes e pela comparabilidade com a literatura; como o equilíbrio não se sustenta em cenários reais, métricas por classe complementam a análise (Seções 4.2 e 4.3). Relatam-se intervalos de confiança de 95% por *bootstrap* de percentil ($B = 10\,000$); teste exato de McNemar par a par com correção de Holm; validação cruzada estratificada de 5 folds; e calibração via *Expected Calibration Error* (ECE) e Brier.

### 3.3. Diagnósticos de generalização

Os diagnósticos são sete: (i) análise lexical por log-odds com *prior* de Dirichlet, sobre o corpus completo, em uso descritivo [Monroe et al. 2008]; (ii) testes de estresse (particionamentos por fonte, temporal e antiviés preliminar categoria $\times$ período); (iii) avaliação cross-dataset bidirecional, com os rótulos dos corpora alinhados a uma convenção única antes dos experimentos e verificação automática de sanidade de polaridade (detalhada adiante); (iv) robustez adversarial por perturbações determinísticas e *back-translation* PT$\rightarrow$EN$\rightarrow$PT [Tiedemann and Thottingal 2020]; (v) equalização de comprimento por sumarização extrativa TextRank adicional da classe verdadeira [Mihalcea and Tarau 2004]; (vi) XAI por LIME sobre o ensemble, com estabilidade medida pelo coeficiente de Jaccard dos *top-10* tokens em execuções repetidas [Ribeiro et al. 2016]; e (vii) *sondas de atalho*: Regressões Logísticas usando apenas o número de tokens, apenas os $K$ termos mais discriminativos por classe (log-odds no

treino) ou apenas metadados (categoria e ano), medindo quanto do desempenho é recuperável sem conteúdo semântico.

Três componentes do protocolo merecem detalhe adicional, por sustentarem a leitura cautelosa dos resultados. *(a) Alinhamento de rótulos.* As convenções numéricas dos corpora divergem (no FakeRecogna 2.0, o rótulo 0 é a classe verdadeira; o Fake.br usa diretórios *fake*/*true*); os rótulos externos são alinhados à convenção do treino *antes* de qualquer predição, com verificação automática de sanidade (acurácia direta vs. inversão hipotética) auditando o alinhamento, sem papel nas métricas. *(b) Particionamento antiviés.* Os estratos combinam *simultaneamente* rótulo, categoria (dez temas mais frequentes; demais em *outros*) e período (quartis do ano de publicação), com fusão de estratos raros, igualando a distribuição desses fatores entre as partições, sem eliminar as correlações entre classe e tema ou período. *(c) Back-translation.* Dois modelos MarianMT (PT→EN→PT) com busca em feixe, sobre um subconjunto de 300 exemplos, medindo a variação de F1-macro frente ao texto original.

## 4. Resultados e Discussão

### 4.1. Desempenho dentro da distribuição (*in-distribution*, IID)

A Tabela 2 apresenta F1-macro, IC de 95% (bootstrap) e ECE das dezessete configurações. O melhor desempenho é do BERTimbau ajustado (F1 = 0,9621; média multi-semente 0,9623, DP 0,0003), seguido pelos ensembles (0,9611–0,9612) e pelo BERTimbau-large (0,9597). O melhor baseline lexical é o TF-IDF+MLP (0,9412); a diferença para o melhor modelo é de apenas 0,0209. A curva de aprendizado satura cedo: com 20% dos dados de treino o ensemble (execução única, semente 42) já atinge F1 = 0,9616, subindo a apenas 0,9628 com 100% dos dados (variação inferior a 0,002, comparável ao desvio entre sementes, $\leq 0{,}0024$).

**Tabela 2. Desempenho IID na variante abstrativa do FakeRecogna 2.0 (teste, $N = 10\,471$): F1-macro com IC 95% por bootstrap ($B = 10\,000$) e ECE, ordenado por F1 (convenção de sementes na Seção 3.2).**

| Modelo | F1-macro [IC 95%] | ECE |
|---|---|---|
| BERTimbau-base (ajustado) | 0,9621 [0,9584–0,9657] | 0,0351 |
| Ens3 (CNN+LSTM+ConvLSTM) | 0,9612 [0,9575–0,9648] | 0,0566 |
| Ens2 (CNN+LSTM) | 0,9611 [0,9574–0,9648] | 0,0568 |
| WEns2 ($\alpha = 0{,}50$) | 0,9611 [0,9574–0,9648] | 0,0568 |
| WEns3 (0,5/0,2/0,3) | 0,9611 [0,9574–0,9648] | 0,0573 |
| LSTM | 0,9598 [0,9560–0,9635] | 0,0515 |
| BERTimbau-large (PLM) | 0,9597 [0,9560–0,9634] | 0,0361 |
| CNN | 0,9591 [0,9554–0,9629] | 0,0540 |
| ConvLSTM | 0,9580 [0,9541–0,9617] | 0,0483 |
| XLM-RoBERTa-base (PLM) | 0,9443 [0,9398–0,9486] | 0,0377 |
| TF-IDF+MLP | 0,9412 [0,9367–0,9456] | 0,0069 |
| TF-IDF+SVM | 0,9402 [0,9356–0,9447] | 0,0057 |
| BERT[CLS]+MLP | 0,9359 [0,9310–0,9405] | 0,0217 |
| TF-IDF+LogReg | 0,9315 [0,9267–0,9363] | 0,0595 |
| BERT[CLS]+SVM | 0,9265 [0,9216–0,9313] | 0,0086 |
| BERT[CLS]+LogReg | 0,9243 [0,9192–0,9292] | 0,0058 |
| mDeBERTa-v3-base (PLM) | 0,9197 [0,9144–0,9248] | 0,0275 |

O teste exato de McNemar par a par com correção de Holm, aplicado à família fixa de sete configurações do protocolo, mostra que todas as redes profundas, ensembles e o BERTimbau ajustado superam o TF-IDF+MLP de forma significativa ($p_{\text{Holm}} < 0{,}001$ em todos os pares contra o baseline). Contudo, *nenhuma* diferença entre as seis demais configurações da família resiste à correção: o menor $p$-valor corrigido entre elas é 0,10 (Ens3 vs. ConvLSTM), e a maioria dos pares satura em $p_{\text{Holm}} = 1{,}0$. A superioridade estatisticamente sustentada restringe-se, portanto, à comparação *com o baseline lexical*; o "topo" é indistinguível. A validação cruzada de 5 folds corrobora a estabilidade (F1-macro 0,9603, DP 0,0038).

### 4.2. Avaliação cross-dataset (fora da distribuição)

A Tabela 3 e a Figura 2 sintetizam a avaliação cross-dataset. Na direção FakeRecogna 2.0 $\rightarrow$ Fake.br, o F1-macro cai de 0,9612 para 0,6959 no Ens3 ($\Delta$F1 = 0,2653) e de 0,9621 para 0,6833 no BERTimbau ajustado ($\Delta$F1 = 0,2788), com acurácias de 0,7000 e 0,6915. A queda é assimétrica por classe: o recall da classe fake permanece alto (0,816 no Ens3; 0,853 no BERTimbau), mas 41,6% e 47,0% das notícias *verdadeiras* são classificadas como fake: fora do corpus de treino, superprevê-se a classe fake. A direção inversa (Fake.br $\rightarrow$ FakeRecogna 2.0) recai na mesma faixa: F1 = 0,6883 no Ens2 (acurácia 0,6887) e F1 = 0,6772 no BERTimbau ajustado (acurácia 0,6800). A simetria entre as duas direções (F1 entre 0,68 e 0,70) indica que o sinal aprendido em cada corpus é, em larga medida, específico daquele corpus. A transferência interna para a variante extrativa não se degrada (BERTimbau 0,9829; Ens3 0,9782), mas trata-se de transferência *intra-corpus* (mesma fonte, sumarização distinta), e não de validação externa.

**Tabela 3. Avaliação cross-dataset (rótulos dos corpora alinhados a uma convenção única antes dos experimentos; a verificação de sanidade de polaridade não acusou desalinhamento).**

| Cenário | Modelo | Acurácia | F1-macro |
|---|---|---|---|
| IID (referência) | BERTimbau ajustado | 0,9621 | 0,9621 |
| IID (referência) | Ens3 | 0,9612 | 0,9612 |
| FakeRecogna $\rightarrow$ Fake.br | BERTimbau ajustado | 0,6915 | 0,6833 |
| FakeRecogna $\rightarrow$ Fake.br | Ens3 | 0,7000 | 0,6959 |
| Fake.br $\rightarrow$ FakeRecogna | BERTimbau ajustado | 0,6800 | 0,6772 |
| Fake.br $\rightarrow$ FakeRecogna | Ens2 | 0,6887 | 0,6883 |
| $\rightarrow$ Extrativa (interna) | BERTimbau ajustado | 0,9829 | 0,9829 |
| $\rightarrow$ Extrativa (interna) | Ens3 | 0,9782 | 0,9782 |

### 4.3. Diagnóstico de regularidades e atalhos

Nos testes de estresse (Tabela 4), o particionamento por fonte eleva a acurácia (0,9770), mas *reduz* o F1-macro (0,9441), com queda concentrada na classe verdadeira (F1 = 0,9013 vs. 0,9870). Os particionamentos temporal (F1 = 0,9803) e antiviés preliminar categoria $\times$ período (F1 = 0,9854) *aumentam* o F1 frente ao particionamento aleatório (0,9611): em vez de indicarem maior generalização, são compatíveis com confundimento temporal e temático.

A análise lexical por log-odds (Tabela 5) revela forte separabilidade lexical alinhada a tema e a gênero de circulação digital: a classe verdadeira concentra termos insti-

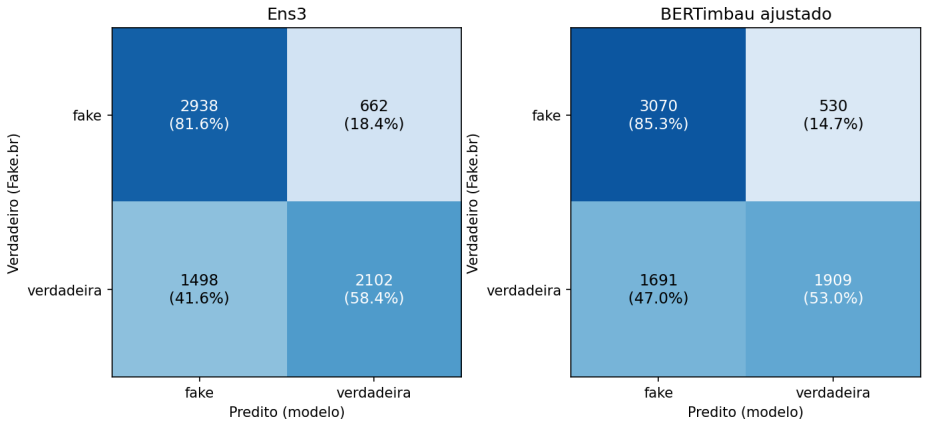



**Figura 2. Matrizes de confusão da avaliação cross-dataset (FakeRecogna 2.0 → Fake.br); contagem e % por linha (classe verdadeira do Fake.br). A transferência é assimétrica: recall alto na classe fake, mas 41,6% (Ens3) e 47,0% (BERTimbau) das notícias verdadeiras classificadas como fake.**

**Tabela 4. Testes de estresse de generalização (variante abstrativa). O particionamento aleatório serve de referência (ensemble CNN+LSTM, F1-macro = 0,9611). O particionamento antiviés é preliminar.**

| Particionamento | Acurácia | F1-macro |
|---|---|---|
| Aleatório (referência) | 0,9611 | 0,9611 |
| Por fonte (*out-of-source*) | 0,9770 | 0,9441 |
| Temporal | 0,9857 | 0,9803 |
| Antiviés categoria×período (prelim.) | 0,9854 | 0,9854 |

tucionais/sanitários (*saúde*, *covid-19*, *pandemia*) e a classe falsa concentra marcadores de circulação em redes (*vídeo*, *redes*, *sociais*, *whatsapp*). O trigrama "circula redes sociais" aparece 2.216 vezes na classe falsa contra 24 na verdadeira. A distribuição temporal por classe reforça o confundimento: a classe verdadeira concentra-se em 2020–2021 (pico da COVID-19), enquanto a classe falsa se distribui ao longo de todo o período (colunas $N_v/N_f$ da Tabela 6). Estratificando o desempenho do Ens3 por ano de publicação (Tabela 6), a acurácia mantém-se alta em todo o período, mas o F1-macro só se sustenta em 2020–2022, os únicos anos em que as duas classes coexistem em volume; nos demais, com uma das classes ausente ou quase ausente (colunas $N_v/N_f$), o colapso do F1-macro é em grande parte um efeito mecânico do suporte. O ponto diagnóstico não é o F1 baixo em si, e sim a distribuição que o produz: rótulo, período e tema estão fortemente correlacionados. A assimetria de comprimento é compatível com um atalho adicional: na faixa curta (0–24 tokens, $N = 2\,586$), a acurácia do Ens3 permanece alta (0,9571), mas o F1-macro cai para 0,8066, com a queda concentrada em uma das classes.

As sondas de atalho tornam essas regularidades explícitas (Tabela 7): apenas o número de tokens atinge F1-macro de 0,7128; apenas os 20 termos mais discriminativos por classe, 0,8445; e apenas metadados (categoria e ano), 0,9611, numericamente equivalente ao ensemble neural (0,9612), sem acesso a uma única palavra do texto: no particionamento aleatório deste corpus, categoria e período praticamente determinam a classe.

**Tabela 5. Termos de conteúdo mais discriminativos por classe (log-odds com *prior* de Dirichlet; estatística $z$); verbos de elocução e termos funcionais de alto $z$ (p. ex., *disse*, *diz*) foram omitidos. Magnitudes maiores indicam associação mais forte.**

| Classe verdadeira | | Classe falsa | |
|---|---|---|---|
| Termo | $z$ | Termo | $z$ |
| saúde | 55,4 | vídeo | −60,0 |
| coronavírus | 48,2 | redes | −59,7 |
| covid-19 | 46,4 | sociais | −56,9 |
| pandemia | 41,6 | lula | −55,4 |
| ministério | 39,2 | facebook | −50,4 |
| mortes | 30,6 | whatsapp | −36,2 |

**Tabela 6. Ens3 no teste IID por ano de publicação ($N = 7.231$ com data), com suporte por classe ($N_v$: verdadeiras; $N_f$: fake). Fora de 2020–2022, com uma classe quase ausente, o colapso do F1-macro é em grande parte mecânico; essa concentração das classes em períodos distintos é o próprio confundimento temporal.**

| Ano | $N$ | $N_v$ | $N_f$ | Acurácia | F1-macro |
|---|---|---|---|---|---|
| 2014–2018 | 832 | 0 | 832 | 0,9639 | 0,4908 |
| 2019 | 495 | 6 | 489 | 0,9697 | 0,6662 |
| 2020 | 2.584 | 1.962 | 622 | 0,9578 | 0,9437 |
| 2021 | 2.307 | 1.803 | 504 | 0,9480 | 0,9276 |
| 2022 | 663 | 30 | 633 | 0,9502 | 0,8053 |
| 2023 | 350 | 0 | 350 | 0,9686 | 0,4920 |

**Tabela 7. Sondas de atalho (Regressão Logística, split oficial, apenas sinais superficiais): F1-macro no teste; última linha: referência.**

| Sonda | Features | F1-macro |
|---|---|---|
| Nº de tokens | 1 | 0,7128 |
| Top-6 termos por classe (log-odds) | 12 | 0,7816 |
| Top-20 termos por classe (log-odds) | 40 | 0,8445 |
| Metadados (categoria + ano) | 12 | 0,9611 |
| Referência: Ens3 (texto completo) | – | 0,9612 |

Essas regularidades remetem ao viés de seleção dos corpora: a classe falsa provém de agências de checagem (conteúdo viral de redes, metatexto de checagem) e a verdadeira de veículos jornalísticos (pauta institucional e sanitária do período coletado), com pareamento e sumarização assimétrica adicionando artefatos próprios. Parte da separabilidade é fabricada pelo processo de amostragem, não pela natureza da desinformação; são esses os sinais que as sondas capturam e que um modelo de maior capacidade pode explorar como atalho.

Quanto à robustez, perturbações superficiais (*typos*, deleção e troca) têm efeito pequeno ($\Delta$F1 $<$ 0,015); a *back-translation* é mais agressiva ($\Delta$F1 = 0,0557 no Ens3; 0,0722 no BERTimbau, em 300 exemplos). A equalização de comprimento (sumarização extrativa adicional da classe verdadeira, aproximando os comprimentos entre as classes) praticamente não altera o desempenho ($\Delta$F1 = −0,0011): o excedente de comprimento,

isolado, não explica o sinal; equalizar a sumarização da classe falsa fica como etapa futura.

Na análise de explicabilidade (XAI), a estabilidade do LIME é limitada: o Jaccard médio dos *top-10* tokens entre execuções é 0,329. Qualitativamente, em três exemplos por célula da matriz de confusão IID, os acertos (TP/TN) tendem a priorizar termos plausíveis (*checagem*, *justiça*, *investigado*), enquanto os erros (FP/FN) priorizam termos metatextuais ou genéricos (*via*, *assim*, *apenas*). Esse padrão é *indício* exploratório (poucos exemplos por célula), não prova de dependência causal. Uma limitação do LIME padrão é operar sobre tokens isolados, pouco informativos quando funcionais; a separabilidade em n-gramas é coberta pela análise lexical (log-odds de uni a trigramas), e explicações sobre n-gramas ficam como extensão.

### 4.4. Discussão

Os resultados sustentam uma leitura cautelosa. O alto desempenho IID ($\sim$0,96) é real e estatisticamente robusto *frente a baselines lexicais*, mas: (i) as melhores arquiteturas são indistinguíveis entre si; (ii) a saturação precoce da curva de aprendizado, a forte separabilidade lexical e as sondas de atalho (metadados sozinhos igualam o ensemble) indicam componente superficial expressivo; (iii) a queda cross-dataset de 27–28 pontos, com F1 entre 0,68 e 0,70 nas duas direções, limita a generalização externa; e (iv) equalização e estresse não isolam uma única causa. A convergência dessas evidências é *compatível* com aprendizado por atalho, mas não comprova o mecanismo decisório; mantém-se a interpretação como hipótese diagnóstica triangulada.

Do conjunto de evidências, extraem-se recomendações práticas para a avaliação de detecção de *fake news* em português, que constituem a face aplicada da contribuição metodológica deste trabalho. Primeiro, métricas intra-corpus por particionamento aleatório são insuficientes: a avaliação cross-dataset deveria ser requisito, não complemento. Segundo, recomenda-se auditar a separabilidade lexical (por exemplo, por log-odds com *prior* de Dirichlet), a distribuição de classes por tema e período e o desempenho de sondas triviais (comprimento, vocabulário restrito, metadados), que expõem confundimentos que as métricas agregadas ocultam. Terceiro, a comparação entre arquiteturas deve apoiar-se em testes pareados com correção para múltiplas comparações e em intervalos de confiança, evitando conclusões a partir de diferenças não significativas. Quarto, a calibração (ECE, Brier) deve ser reportada, sobretudo *fora da distribuição*, onde a degradação da incerteza é esperada sob *shift* [Quiñonero-Candela et al. 2009]. Em conjunto, essas práticas deslocam o foco de *quanto* um modelo acerta para *por que* acerta.

Quanto às implicações práticas, sob a suposição plausível de que em fluxos operacionais o conteúdo legítimo é majoritário e o custo de rotular notícia verdadeira como falsa é alto, a precisão da classe fake passa a dominar a utilidade do sistema; no cenário cross-dataset ela cai de 0,96–0,97 (IID) para 0,64–0,66, e os alertas passariam a ser em sua maioria incorretos em fluxos com menos de cerca de um terço de conteúdo falso. Modelos com esse perfil devem operar como apoio à triagem de checadores humanos (ordenação de fila, não veredito automático), com monitoramento de *drift* e auditoria de atalhos antes de reuso fora do domínio de treino.

As limitações e ameaças à validade são de *construto* (“atalho” é hipótese diagnóstica, não causa comprovada; e os critérios de rotulagem dos corpora diferem: desmentidos de checagem vs. notícias falsas na íntegra), *externa* (avaliação principal em um

único corpus, com transferência cross-dataset frágil) e *de medida* (redes profundas relatadas como média de probabilidades entre sementes, não execução única; e as linhas direta e de transferência interna do BERTimbau ajustado derivam de execução anterior ao isolamento dos artefatos multi-semente, corroboradas em magnitude pelas linhas reexecutadas de forma independente).

## 5. Conclusão e Trabalhos Futuros

Mostrou-se que, em detecção de *fake news* em português brasileiro, alto desempenho dentro da distribuição coexiste com generalização frágil: das dezessete configurações avaliadas no FakeRecogna 2.0, as melhores atingem F1-macro $\sim$0,96, com topo estatisticamente indistinguível, mas a generalização cross-dataset colapsa: queda de 27–28 pontos em F1-macro, com F1 entre 0,68 e 0,70 nas duas direções. Análises lexicais, de curva de aprendizado, de estresse, adversariais, de XAI e as sondas de atalho diagnosticam regularidades do corpus compatíveis com aprendizado por atalho (hipótese, não prova causal). A contribuição metodológica é um *protocolo crítico* (IC por bootstrap, McNemar com Holm, avaliação cross-dataset, auditoria lexical por log-odds, equalizadores, sondas de atalho e XAI diagnóstica), sintetizado na Figura 1 e recomendado para a tarefa.

Como trabalhos futuros, planeja-se: a triangulação quantitativa de XAI por gradiente (Integrated Gradients e Attention Rollout) sobre os mesmos exemplos, considerando o *disagreement problem*, a baixa concordância entre métodos de explicabilidade [Krishna et al. 2022]; uma auditoria humana de erros com múltiplos avaliadores; um relatório formal de deduplicação e vazamento; o particionamento antiviés final; e a extensão da avaliação cross-dataset a corpora adicionais (p. ex., FACTCK.BR e FakeTrueBR), condicionada à compatibilidade de rótulos e licenças.

## Agradecimentos

Os autores agradecem à FAPESP (processos 2019/07665-4 e 2024/15527-9) e ao CNPq (processo 309525/2026-8) pelo apoio a este projeto. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

## Referências

Abonizio, H. Q., de Morais, J. I., Tavares, G. M., and Barbon Junior, S. (2020). Language-independent fake news detection: English, Portuguese, and Spanish mutual features. *Future Internet*, 12(5):87.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics.

Fischer, M., Haque, R., Stynes, P., and Pathak, P. (2022). Identifying fake news in Brazilian Portuguese. In *Natural Language Processing and Information Systems (NLDB 2022)*, volume 13286 of *Lecture Notes in Computer Science*, pages 110–122. Springer.

Garcia, G. L., Paiola, P. H., Jodas, D. S., Sugi, L. A., and Papa, J. P. (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 (PROPOR 2024)*, volume 1, pages 86–96. Association for Computational Linguistics.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Guedes, R., Barros, B., and do Nascimento, H. (2026). Identificação de notícias falsas em português: um olhar sobre a generalização de modelos. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*, volume 1, pages 717–726. Association for Computational Linguistics.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 2, pages 107–112. Association for Computational Linguistics.

He, P., Gao, J., and Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543.

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

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, Z. S., and Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv:2202.01602.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411. Association for Computational Linguistics.

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in Portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language (PROPOR 2018)*, volume 11122 of *Lecture Notes in Computer Science*, pages 324–334. Springer.

Okano, E. Y., Liu, Z., Ji, D., and Ruiz, E. E. S. (2020). Fake news detection on Fake.Br using hierarchical attention networks. In *Computational Processing of the Portuguese Language (PROPOR 2020)*, volume 12037 of *Lecture Notes in Computer Science*, pages 143–152. Springer.

Oshikawa, R., Qian, J., and Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and*

*Evaluation Conference (LREC)*, pages 6086–6093. European Language Resources Association.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. MIT Press.

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 (KDD)*, pages 1135–1144. ACM.

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.

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

Soares, F. B., Recuero, R., Volcan, T., Fagundes, G., and Sodré, G. (2021). Research note: Bolsonaro's firehose: How Covid-19 disinformation on WhatsApp was used to fight a government political crisis in Brazil. *Harvard Kennedy School Misinformation Review*, 2(1):1–17.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems (BRACIS 2020)*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, pages 479–480.

Vicentini, J., Rodrigues, R. B. d. M., Candido Junior, A., and Guilherme, I. R. (2026). Comparing explainable AI techniques in language models: A case study for fake news detection in Portuguese. *Journal of the Brazilian Computer Society*, 32(1):1–12.

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

Wanderley, M. G., Ferraz, L. B. S., de Almeida, T. A., and Silva, R. M. (2025). A moving target: Detecting concept drift in Brazilian Portuguese fake news. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 490–501. SBC.

Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Technical Report DGI(2017)09, Council of Europe.