# Identificação de notícias falsas em português: um olhar sobre a generalização de modelos

**Raphael Guedes[1], Bruno Barros[1], Hugo do Nascimento[1]**
[1]Instituto de Informática, Universidade Federal de Goiás, Goiás, Brasil,
{raphaelguedes, hadn}@ufg.br, brunomattos@discente.ufg.br

## Resumo

A disseminação de desinformação em meios digitais requer mecanismos robustos de detecção, tarefa na qual modelos de linguagem apresentam desempenho satisfatório. Entretanto, são percebidas na literatura análises que desconsideram a característica da degradação da capacidade de generalização dos modelos em dados reais, diferentes daqueles nos quais o treino ou ajuste fino foi realizado. Este trabalho investiga o comportamento dos modelos BERTimbau e mBERT em cenários de generalização cruzada (dados de teste diferentes dos dados de treinamento e validação). Para isso, foi realizado um ajuste fino utilizando quatro *corpora* brasileiros (Fake.br, Fakepedia, FakeRecogna e FakeTrueBR). Os resultados confirmam a hipótese de que avaliações intra-base têm altas taxas de desempenho, enquanto avaliações entre-bases têm baixas taxas e alta degradação na generalização cruzada, ainda que o objetivo de identificação de notícias falsas seja mantido. Quanto à capacidade preditiva dos modelos, o BERTimbau se mostrou ligeiramente melhor na média com 71% de acurácia e 67% de f1-score contra 69% e 64%, respectivamente, para o mBERT.

## 1 Introdução

A alta geração de informações nos meios digitais, principalmente por meio das plataformas de redes sociais, é um fenômeno presente na sociedade contemporânea. No entanto, em meio à vastidão de textos, imagens, vídeos e áudios produzidos, são encontrados tanto conteúdos verídicos quanto inverídicos. As *fake news* são definidas como alegações falsas ou enganosas que imitam o estilo de notícias reais (Gelfert, 2018; Monteiro et al., 2018; Baracho et al., 2025), apresentando diferentes categorizações quanto à sua natureza e intensidade (Rubin et al., 2015; Wardle e Derakhshan, 2017). Os danos causados pela sua disseminação ecoam por diferentes esferas da sociedade. No contexto brasileiro, as eleições de 2018 e a pandemia da COVID-19 são exemplos de momentos marcados pelo alto volume de conteúdo falso compartilhado em massa, conforme se verifica em Forte Martins et al. (2021); Charles et al. (2022) e Moreira et al. (2023).

Nesse cenário, os modelos de linguagem contextuais surgem como ferramentas fundamentais no combate à desinformação (Moreira et al., 2023; Pires e Guerreiro, 2024). Devido à capacidade de processar grandes volumes de dados e apreender aspectos linguísticos, essas técnicas permitem diferenciar conteúdos reais de falsos com alta precisão. Todavia, bases de notícias carregam o contexto temporal e social do período em que foram construídas e, segundo Baracho et al. (2025), os *corpora* existentes para o português brasileiro possuem baixa variabilidade temática. Isso configura uma limitação que pode provocar degradação na capacidade de generalização preditiva dos modelos. Tal fenômeno consiste na perda da capacidade de classificação de novas entradas que pertençam a outro conjunto de dados, ainda que tratem do mesmo tema, como pontuado por Cabral et al. (2021) e Avram et al. (2025). Além disso, observa-se nos trabalhos a tendência de avaliar a identificação de *fake news* com ênfase em uma perspectiva intra-base (treino e teste no mesmo conjunto de dados), negligenciando cenários entre-bases (treino em um conjunto, teste em outro).

Diante disso, este trabalho busca responder à seguinte questão: "Como se comportam modelos de linguagem pré-treinados em português brasileiro quando finalmente ajustados a um conjunto de dados e aplicados a conjuntos distintos, para a detecção de notícias falsas?". É considerada a hipótese de que modelos baseados em *Transformers* (*Bidirectional Encoder Representations from Transformers* – BERT) apresentam degradação na capacidade de generalização, quando avaliados em dados diferentes daqueles originalmente treinados.



*Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*
*Vol. 1*, pages 717–726

Assim, o presente trabalho objetiva avaliar a eficácia de variações do BERT na identificação de notícias falsas em português, focando na sua capacidade de generalizar conhecimento entre diferentes conjuntos de dados. Como objetivos específicos, almeja-se: (1) realizar um ajuste fino (*fine-tuning*) dos modelos BERTimbau [1] e *Multilingual* BERT (mBERT) [2] em um conjunto de dados de referência e avaliar a capacidade de generalização de conhecimento em conjuntos desconhecidos pelos modelos; (2) comparar o desempenho de cada modelo para determinar aquele com maior robustez na generalização.

Destaca-se, na literatura, o trabalho de Pires e Guerreiro (2024) que, ainda que faça intersecção no uso dos modelos e bases de dados desta pesquisa, buscou avaliar se uma instância do modelo BERT pré-treinada em português é mais eficaz se comparada ao modelo original e ao mBERT. O foco residiu em comparar as versões do modelo e não no fenômeno da degradação dos resultados. Ademais, a validação dos modelos ocorreu em um contexto intra-base.

O restante do trabalho está organizado da seguinte forma: a Seção 2 estabelece a base teórica e os principais achados da literatura; a Seção 3 descreve as bases de dados, os modelos utilizados e o protocolo experimental; a Seção 4 apresenta e discute os resultados obtidos; por fim, a Seção 5 conclui o trabalho tecendo considerações sobre as limitações deste estudo e apresentando possibilidades para investigações futuras.

## 2 Trabalhos Relacionados

### 2.1 Bases de dados de notícias falsas em português

Segundo o relatório apresentado por Meltwater (2025), o Brasil é a quinta maior população conectada do mundo, com cerca de 86% de seus habitantes com acesso à *internet*. Entretanto, Baracho et al. (2025), ao analisarem bases de dados de notícias falsas em português brasileiro, apontam que a pesquisa científica neste idioma ainda carece de ampliação, o que acaba limitando o avanço tecnológico e o desenvolvimento de soluções de Processamento de Linguagem Natural (PLN) para o contexto nacional.

Considerado o primeiro *corpus* construído para o idioma, o *Fake.br* foi proposto por Monteiro et al. (2018) para suprir a necessidade de dados anotados. O processo consistiu em coletar e verificar manualmente as notícias falsas, seguido por uma etapa semiautomática de coleta com similaridade lexical baseada em cosseno para garantir o pareamento temático. Eles evidenciaram ser importante o alinhamento entre o falso e o verdadeiro, aplicado também por outros autores adiante, o que permite a validação de padrões linguísticos nos dados.

Posteriormente, outros dois *corpora* foram desenvolvidos: o *FakeRecogna* (Garcia et al., 2022) e o *Fakepedia* (Charles et al., 2022). O *FakeRecogna*, maior e mais recente que o *Fake.br*, não passou por alinhamento dos dados, somente por balanceamento entre as classes. Seu processo de coleta utilizou *crawlers* e a base foi avaliada em um conjunto de modelos de aprendizado de máquina, após ser vetorizada com *Bag of Words* (BoW) e *FastText*. Por sua vez, o *Fakepedia* foi projetado para ser flexível e facilmente atualizável. A coleta das notícias falsas para essa base também foi feita com *crawlers*, enquanto as verdadeiras foram extraídas a partir de palavras-chave presentes nas falsas. O alinhamento foi realizado selecionando a notícia mais similar dentre cinco candidatas e utilizou o algoritmo Word2Vec (Mikolov et al., 2013) para cálculo de similaridade de cosseno.

No entanto, o trabalho de Chavarro et al. (2023) destaca falhas no alinhamento do *Fakepedia*, propondo, por outro lado, a criação de um novo *corpus*, o *FakeTrue.br*. Ainda que seja o menor conjunto dentre os apresentados nesta subseção, seu destaque se deve ao processo de alinhamento, alcançado com a utilização de vetores (*embeddings*) gerados pelo Sentence-BERT multilíngue (Reimers e Gurevych, 2019). Para eles, o modelo de linguagem permitiu calcular, mais precisamente, a similaridade semântica, corrigindo problemas de baixa semelhança observados em outras bases.

Uma particularidade percebida pelos autores do presente estudo, bem como por Baracho et al. (2025) é a recorrência de uso das mesmas fontes de informação, tanto para os dados reais como para os inverídicos (como mostra a Tabela 1). Esses autores, juntamente com de Morais et al. (2020), apontam que o uso das mesmas origens aumenta a probabilidade de enviesamento dos dados, pois cada agência possui critérios editoriais específicos.

Embora a criação desses *corpora* represente um avanço significativo, Baracho et al. (2025) informam que ainda são necessários mais esforços para alcançar o patamar dos recursos disponíveis para

[1] Ver Subseção 2.2.
[2] Ver Subseção 2.2.

outras línguas, como o inglês, que possui bases com milhões de textos. Outro ponto crítico é a pouca variabilidade temática nas bases em português, o que, se resolvido, constitui aspecto importante para garantir boa generalização dos modelos na identificação de notícias falsas em diferentes domínios.

### 2.2 BERT e suas variações

O *Bidirectional Encoder Representations from Transformers* (BERT) foi proposto por Devlin et al. (2019). De acordo com eles, o modelo, baseado na arquitetura *Transformer Encoder*, foi desenvolvido para aprender representações bidirecionais a partir de texto não rotulado. O seu pré-treinamento emprega duas abordagens não supervisionadas: *Masked Language Model* (MLM), que mascara aleatoriamente uma porcentagem dos *tokens* de entrada exigindo que o modelo prediga os termos a partir do contexto, e *Next Sentence Prediction* (NSP), que lhe possibilita compreender a relação entre pares de frases.

O *Multilingual* BERT (mBERT) é uma variante do BERT treinada em 104 idiomas com conteúdo proveniente da *Wikipédia*[3] e com um vocabulário com cerca de 120.000 *tokens*. Apesar de possuir a habilidade de compreensão de várias línguas, Souza et al. (2020) alertam que essa abrangência pode prejudicar a eficácia preditiva em línguas específicas, um fenômeno conhecido como "maldição da multilingualidade".

O *BERT for Brazilian Portuguese* (BERTimbau), apresentado no estudo de Souza et al. (2020), é um modelo em versões base e grande (*large*) pré-treinado no *corpus Brazilian Web as Corpus* (brWaC) (Wagner Filho et al., 2018), totalmente em português brasileiro. Avaliações realizadas revelaram que ele apresenta melhor desempenho que o mBERT (Pires e Guerreiro, 2024) e outros modelos multilíngues em tarefas de PLN para o português, beneficiando-se de um vocabulário adaptado às características do idioma, e possibilita uma segmentação de palavras mais apropriada.

### 2.3 Identificação de notícias falsas em português

O estudo sobre a identificação de notícias falsas em português é recente, com os primeiros trabalhos coletados para a construção desta investigação datando de 2018.

Os trabalhos mais antigos explorados, bem como aqueles que definem as bases de dados aqui empregadas, concentram as análises preditivas em abordagens supervisionadas clássicas, empregando técnicas de vetorização como *Bag of Words* e Word2Vec. O trabalho de Monteiro et al. (2018) é um exemplo. Ao proporem o *Fake.br*, utilizaram Máquinas de Vetores de Suporte (SVM), que demonstraram desempenho robusto na classificação binária. Adicionalmente, de Morais et al. (2020) investigaram a classificação multirrótulo utilizando *Random Forest* e características estilométricas extraídas dos textos, mostrando a eficácia de métodos de *ensemble* ao lidar com as nuances textuais.

Com a evolução das técnicas de PLN, abordagens baseadas em Aprendizado Profundo (AP) ganharam destaque, explorando representações vetoriais densas e redes neurais. A pesquisa de Paixão et al. (2020) propôs um método utilizando Redes Neurais Convolucionais (CNN) alimentadas por vetores de palavras (*word embeddings*), superando classificadores tradicionais com f1-score de 96%. Na mesma linha, Moreira et al. (2021) avaliaram o uso de *Long Short-Term Memory* (LSTM) combinado a representações vetoriais, demonstrando que as arquiteturas que capturam dependências temporais em textos longos oferecem resultados mais promissores para a língua portuguesa.

A introdução de modelos baseados na arquitetura *Transformer* marcou um grande avanço na área, permitindo a captura de contextos bidirecionais complexos. Ao compararem o desempenho do BERTimbau contra algoritmos clássicos e o BERT original, Moreira et al. (2023) confirmaram o desempenho superior do modelo pré-treinado em português. Corroborando esses achados, a análise comparativa de Pires e Guerreiro (2024) entre BERT, mBERT e BERTimbau em múltiplos conjuntos de dados concluiu que o BERTimbau apresenta melhores métricas de acurácia e f1-score, evidenciando a importância do pré-treinamento no idioma alvo para a tarefa de classificação de desinformação.

Recentemente, as pesquisas têm abordado desafios tais como a generalização entre diferentes domínios e a detecção de conteúdo gerado por Grandes Modelos de Linguagem, *Large Language Models*, (LLMs). Garcia et al. (2024) propuseram o uso de sumarização automática (extrativa e abstrativa) combinada a modelos temporais e *Transformers* para lidar com o tamanho dos textos, obtendo alta acurácia com o uso de representações baseadas em *FastText* e *Bidirectional Gated Recurrent Unit* (BiGRU). A pesquisa inovadora de Silva

[3] https://www.wikipedia.org/

et al. (2025) avaliou o impacto de notícias falsas sintéticas geradas por LLMs, em função da robustez de classificadores tradicionais e propondo o uso de *Retrieval-Augmented Generation* (RAG) para mitigar a degradação de desempenho observada quando há incompatibilidade entre as distribuições dos dados de treino, gerados por humanos, e teste, sintéticos.

## 3 Materiais e Métodos

### 3.1 Dados

Os dados utilizados neste estudo são compostos por quatro bases de textos de notícias falsas em português brasileiro: *Fakebr* (Monteiro et al., 2018)[4], *Fakepedia* (Charles et al., 2022)[5], *FakeRecogna* (Garcia et al., 2022)[6] e *FakeTrueBR* (Chavarro et al., 2023)[7]. A Tabela 1 apresenta uma visão geral de cada base.

Embora existam outros conjuntos em português brasileiro, conforme citam Baracho et al. (2025); Cabral et al. (2021); Garcia et al. (2024) e Moreno e Bressan (2019), eles não foram incorporados neste estudo devido à sua especificidade, por apresentarem apenas uma classe ou serem derivados dos quatro primeiros.

Para os dados serem utilizados plenamente nos experimentos, foi preciso padronizar sua estrutura, pois cada base possui uma organização particular. Os dados do *Fake.br* estavam organizados em arquivos individuais para cada classe de notícia (verdadeira ou falsa), e outros para os metadados. Nesse caso, foi feita uma união, gerando um arquivo único. Na base *FakeRecogna*, as colunas `título` e `notícia` foram unidas (gerando o atributo `text`), enquanto o atributo `subtítulo` foi ignorado por apresentar a refutação, em casos de notícias falsas. Já os dados do conjunto *Fakepedia* estavam alinhados em um único arquivo, sendo necessário empilhar as notícias falsas e verdadeiras e unir as suas colunas `title` e `message` (gerando o atributo `text`). Para os dados do *FakeTrueBR*, a organização consistiu no empilhamento dos dados alinhados e na junção dos atributos `title_fake` e `fake` (gerando o atributo `text`).

Ao término da padronização, todos os dados foram armazenados no formato *comma-separated values* (CSV). Os atributos utilizados no ajuste fino e predição foram `text` e `label`. Todos os rótulos (ver Tabela 1) foram mapeados para $1 : fake$ e $0 : true$.

Devido à habilidade do BERT em manipular texto puro em linguagem natural para realizar um aprendizado contextual, apenas a eliminação de quebras de linha, tabulações, espaços em branco extras e linhas em branco foi executada, já que não foram detectados outros tipos de problemas nos dados.

### 3.2 Modelos

Foram utilizadas duas versões do modelo BERT: o mBERT (Devlin et al., 2019) e o BERTimbau (Souza et al., 2020). O mBert usado foi a versão "*bert-base-multilingual-cased*"[8] e o BERTimbau foi o "*neuralmind/bert-large-portuguese-cased*"[9]. Os dois modelos foram aplicados em suas versões capitalizadas (*cased*).

O ajuste fino foi configurado com três épocas, com *batchsize* igual a 8 e taxa de aprendizado (*learning rate*) definida em $2e-5$, seguindo as recomendações de Devlin et al. (2019).

Para a avaliação de desempenho, adotaram-se as métricas padrão da literatura: acurácia para medir a proporção de acertos em relação ao total de dados; precisão para avaliar a proporção de classificações corretas dentre as preditas como positivas (*fake*); revocação (*recall*) para indicar a fração de instâncias reais que foram corretamente identificadas pelo modelo; e f1-score que representa a média harmônica entre a precisão e a revocação.

Os experimentos foram executados em uma máquina equipada com um processador Ryzen 9 7950X, com 16 núcleos e 32 *threads*, 64 GB de memória RAM DDR5 e uma GPU NVIDIA RTX4080 com 16 GB de VRAM.

### 3.3 Abordagem Experimental

A abordagem de experimentação para avaliação da degradação da generalização dos modelos em função da variação dos dados se deu a partir da definição de um protocolo experimental. Seja $D$ o conjunto de todas as bases de dados disponíveis, com $D = \{D_1, D_2, \ldots, D_n\}$ e $n \in \mathbb{N}^*$. Cada base $D_i$ consiste em um conjunto de pares (notícia (*text*), classe (*label*)): $D_i = \{(x_k, y_k)\}_{k=1}^{|D_i|}$, onde $i \leq n$,

[4] https://github.com/roneysco/Fake.br-Corpus

[5] https://github.com/andersoncordeiro/Fakepedia-Corpus

[6] https://github.com/Gabriel-Lino-Garcia/FakeRecogna

[7] https://github.com/jpchav98/FakeTrue.Br

[8] https://huggingface.co/google-bert/bert-base-multilingual-cased

[9] https://huggingface.co/neuralmind/bert-large-portuguese-cased

| | Fake.Br | FakeRecogna | Fakepedia | FakeTrueBR |
|---|---|---|---|---|
| **Qtd. de Registros** | 7.200 | 11.902 | 12.398 | 3.582 |
| **Origem (Fake)** | A Folha do Brasil, Diário do Brasil, The Jornal Brasil, Top Five TV | AFP Checamos, Boatos.org, E-farsas, Fato ou Fake, UOL Confere, Projeto Comprova | Aos Fatos, Boatos.org, Lupa | Boatos.org |
| **Origem (Real)** | G1, Estadão, Folha de São Paulo | G1, Extra, Ministério da Saúde, UOL | G1, Extra, Folha de São Paulo, UOL | G1, Folha de São Paulo |
| **Categorias** | Ciência & Tecnologia, Economia, Política, Religião, Sociedade & Cotidiano, TV & Celebridades | Brasil, Ciência, Entretenimento, Mundo, Política, Saúde | Política, Saúde, Tecnologia, entre outros | Brasil, Entretenimento, Mundo, Política, Saúde |
| **Período de Coleta** | 2016 – 2018 | 2019 – 2021 | 2013 – 2021 | 2017 – 2023 |
| **Classes** | fake, true | 0: fake; 1: true | 0: true, 1: fake | fake, true |
| **Qtd. Atributos** | 28 | 8 | 12 | 5 |
| **Tam. Médio*** | Fake (1.117,9), True (6.671) | Fake (540,9), True (763,3) | Fake (710,7), True (10.915,0) | Fake (932,4), True (3.015,1) |

Tabela 1: Descrição dos conjuntos de dados utilizados. *Tamanho médio de caracteres dos textos

$x_k$ é o texto da notícia $k$ nessa base e $y_k \in \{0, 1\}$ é a sua classe ($0 = verdadeiro$, $1 = falso$).

O experimento foi conduzido iterativamente, onde, a cada iteração, uma base foi escolhida como origem ($D_{src}$) e as restantes como destino. A base de origem $D_{src}$ foi seccionada de acordo com a estratégia *hold-out*, dividindo-a em subconjuntos de treino ($D_{src}^{train}$), validação ($D_{src}^{val}$) e teste ($D_{src}^{test}$), em 80%, 15% e 5%, respectivamente.

Para mensurar a capacidade de generalização, cada modelo ajustado foi testado individualmente nas bases restantes. Em contrapartida ao processo de ajuste fino, a avaliação de degradação utilizou a integralidade dos dados das bases destino, visto que estas são desconhecidas pelo modelo. Portanto, a avaliação dos resultados é dada por $Eval(f(D_i^{train}), D_j)$, para todo $j \in \{1, \ldots, n\}$ tal que $j \neq i$, onde $f$ representa o modelo ajustado e $Eval()$ a função que calcula as métricas definidas.

Essa abordagem permite comparar o desempenho dos modelos no domínio original com os novos, permitindo verificar se os padrões apreendidos em um determinado conjunto $D_{src}$ se mantêm quando aplicados em contextos diferentes.

Por fim, a parte de avaliação dos resultados se deu em três etapas, visando a completude dos objetivos propostos no trabalho: (1) desempenho intra-base, como os dados se comportam na abordagem tradicional de ajuste fino, reproduzindo aquilo encontrado na literatura; (2) análise de generalização entre as bases, que avalia a degradação dos modelos a partir das métricas definidas na Subseção 3.2; e (3) comparação do desempenho das variações do BERT no cenário de degradação da generalização.

| Base | Modelo | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|
| Fake.br | BERTimbau | 0.99 | 0.99 | 0.99 | 0.99 |
| | mBERT | 0.98 | 0.98 | 0.98 | 0.98 |
| Fakepedia | BERTimbau | 1.00 | 1.00 | 1.00 | 1.00 |
| | mBERT | 1.00 | 1.00 | 1.00 | 1.00 |
| FakeRecogna | BERTimbau | 1.00 | 1.00 | 1.00 | 1.00 |
| | mBERT | 1.00 | 1.00 | 1.00 | 1.00 |
| FakeTrueBR | BERTimbau | 1.00 | 1.00 | 1.00 | 1.00 |
| | mBERT | 0.99 | 0.99 | 0.99 | 0.99 |

Tabela 2: Desempenho médio intra-base. Acc.: acurácia, Prec.: precisão, Rec.: recall, e F1: f1-score.

## 4 Resultados e Discussão

Os resultados apresentados nesta seção foram norteados pela questão de pesquisa que buscou investigar o comportamento da aplicação do BERT a bases distintas daquelas utilizadas no treinamento, e validar a hipótese de degradação da capacidade de generalização dos modelos.

### 4.1 Desempenho intra-base

A Tabela 2 evidencia que, ao se adotar a abordagem *hold-out* para segmentação dos dados e ajuste fino, os resultados obtidos são elevados, com diversas ocorrências de 100% de acerto nas métricas. Tais resultados indicam que, quando os dados são avaliados em seu próprio contexto, o desempenho é bastante preciso. Isso é verificado nos trabalhos de Aljawarneh e Swedat (2022) e de Pires e Guerreiro (2024). Ainda que desejáveis, esses valores sugerem a possibilidade de sobreajuste dos modelos.

### 4.2 Degradação do Desempenho

A avaliação dos dados no cenário no qual os modelos finamente ajustados foram testados em bases distintas, confirmou a hipótese do trabalho no que diz respeito à degradação dos modelos, conforme apresentado na Tabela 3.

Ao considerar o *corpus Fake.br*, o desempenho caiu drasticamente quando os modelos foram testados no *FakeRecogna*. O BERTimbau registrou acurácia de 35% e f1-score de 31%, o que o torna inferior a um classificador aleatório. Para o mBERT, a acurácia foi de 34% e o f1-score de 32%. O melhor resultado foi com a base *Fakepedia*, que variou entre 67% e 75%. Ressalta-se que *Fake.br* é a base com a maior variabilidade de temas, porém com dados mais antigos. Todas as outras bases possuem dados mais recentes, como evidencia a Tabela 1, com exceção de *Fakepedia*, a maior em período de coleta.

Quando se muda a análise para o próximo *corpus*: *Fakepedia*, nota-se um desempenho expressivamente superior tanto do BERTimbau quanto do mBERT em relação ao *FakeTrueBR*. As bases compartilham fontes comuns para notícias falsas e reais. É possível que o modelo tenha avaliado os mesmos dados ou muito similares. Nesse bloco de avaliação, *FakeRecogna* é a base cujos resultados mais se degradam em relação a *Fakepedia*. Ao analisar o desempenho médio da base em relação a todas as métricas e modelos, *Fakepedia* é a base que proveu a melhor generalização.

Partindo para a base *FakeRecogna*, constata-se que ela agrupa a maior quantidade de resultados menos expressivos quando analisadas as médias das métricas por base (a saber, acurácia: 63% para BERTimbau; revocação: 52% para BERTimbau e 54% para mBERT; e f1-score: 59% para BERTimbau). Ao retornar à Tabela 3, verifica-se que o conjunto de dados com melhor desempenho foi o *Fakepedia* com 58% de acurácia para o BERTimbau e 62% para o mBERT. Um fator que pode justificar os resultados é a falta de alinhamento entre os textos falsos e reais.

Por fim, o *corpus FakeTrueBR* demonstrou baixa generalização para alguns casos específicos.

Os modelos, quando aplicados ao *Fakepedia*, apresentaram bons resultados, com a ressalva de que as taxas são inferiores para o caso inverso (ajuste fino em *Fakepedia*, aplicado a *FakeTrueBR*).

Após essas análises, observando que alguns conjuntos de dados apresentaram desempenho significativamente superior em relação a outros, decidiu-se investigar se havia sobreposição de dados entre eles. Para isso, foi calculada a sobreposição segundo a expressão $len(A \cap B)/len(A)$, em que $A$ e $B$ são conjuntos de dados distintos e o resultado indica o quanto de $A$ está dentro de $B$. O procedimento ocorreu da seguinte maneira: cada conjunto foi previamente processado, removendo-se espaços em branco, URLs, acentuação e pontuação, além de converter todo o texto para letras minúsculas com uma etapa de lematização. Em seguida, para cada notícia de $A$ que também estivesse presente em $B$ (comparação de igualdade entre sequências de caracteres), o conjunto de intersecção era atualizado e, ao final, obtinha-se a porcentagem desse conjunto em relação ao total de elementos do primeiro. Os resultados estão apresentados na Tabela 5. Ressalta-se que a tabela não é simétrica, pois os valores de interseção são percentuais calculados com base na dimensão de cada conjunto.

Verifica-se que, conforme previamente assumido, os *corpora FakeTrueBR* e *Fakepedia* exibem elevadas taxas de sobreposição de dados. Diante disso, torna-se plausível considerar que o bom desempenho nos testes entre bases decorre desse vazamento de informações, e não de uma generalização efetiva do problema por parte do modelo.

Apesar de não haver sobreposição ao nível de notícias entre as bases *Fakepedia*, *FakeTrueBR* e *Fake.br*, verifica-se que as fontes verídicas se sobrepõem em grande parte dos casos, como apresenta a Tabela 1. Isso pode explicar os resultados superiores nos testes entre essas bases.

### 4.3 Comparação BERTimbau e mBERT

Ao partir para a comparação entre as variações dos modelos, a instância do BERT pré-treinada para o português brasileiro apresenta resultados melhores do que a variação multilíngue, como pode ser constatado na Tabela 4. Isso confirma os achados de Souza et al. (2020) quanto à robustez de um modelo específico para a linguagem. Essa superioridade é verificada em cenários com o ajuste fino feito no conjunto *FakeTrueBR* e a avaliação no *Fake.br*, onde o BERTimbau apresentou 68% de f1score, contra 34% para o mBERT e na generalização entre *Fakepedia* e *Fake.br*, em que obteve 70% de revocação, contra 52% para o modelo pré-treinado em diversos idiomas. Destaca-se, todavia, que a diferença média entre os dois modelos, no contexto de análise da degradação dos resultados, não é muito distante.

| | | Acurácia | | Precision | | Recall | | F1-Score | |
|---|---|---|---|---|---|---|---|---|---|
| **Base de Treino** | **Base de Teste** | *BTb* | *mB* | *BTb* | *mB* | *BTb* | *mB* | *BTb* | *mB* |
| Fake.br | Fake.br | 0.99 | 0.98 | 0.99 | 0.98 | 0.99 | 0.98 | 0.99 | 0.98 |
| | Fakepedia | **0.75** | **0.68** | **0.76** | **0.71** | **0.75** | **0.68** | **0.75** | **0.67** |
| | FakeRecogna | *0.35* | *0.34* | *0.30* | *0.32* | *0.35* | *0.34* | *0.31* | *0.32* |
| | FakeTrueBR | 0.62 | 0.66 | 0.62 | 0.66 | 0.62 | 0.66 | 0.61 | 0.66 |
| Fakepedia | Fake.br | 0.70 | *0.52* | 0.76 | 0.58 | 0.70 | *0.52* | 0.68 | *0.39* |
| | Fakepedia | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | FakeRecogna | *0.49* | 0.53 | *0.44* | *0.58* | *0.49* | 0.53 | *0.35* | 0.42 |
| | FakeTrueBR | **1.00** | **0.99** | **1.00** | **0.99** | **1.00** | **0.99** | **1.00** | **0.99** |
| FakeRecogna | Fake.br | 0.52 | 0.54 | **0.70** | **0.71** | 0.52 | 0.54 | *0.37* | *0.42* |
| | Fakepedia | **0.58** | **0.62** | 0.59 | 0.63 | **0.58** | **0.62** | **0.57** | **0.61** |
| | FakeRecogna | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | FakeTrueBR | *0.44* | *0.50* | *0.43* | *0.50* | *0.44* | *0.50* | 0.42 | 0.50 |
| FakeTrueBR | Fake.br | 0.70 | *0.50* | 0.79 | 0.73 | 0.70 | *0.50* | 0.68 | *0.34* |
| | Fakepedia | **0.76** | **0.65** | **0.84** | **0.75** | **0.76** | **0.66** | **0.75** | **0.62** |
| | FakeRecogna | *0.49* | 0.51 | *0.30* | *0.52* | *0.49* | 0.51 | *0.33* | 0.42 |
| | FakeTrueBR | 1.00 | 0.99 | 1.00 | 0.99 | 1.00 | 0.99 | 1.00 | 0.99 |

Tabela 3: Métricas gerais para toda a base. Em **negrito**, os resultados mais expressivos; em *itálico*, os menos expressivos. mB: mBERT, BTb: BERTimbau

| Modelo | Acurácia | Precision | Recall | F1-Score |
|---|---|---|---|---|
| BERTimbau | **0.71** | 0.72 | **0.71** | **0.67** |
| mBERT | 0.69 | **0.73** | 0.69 | 0.64 |

Tabela 4: Comparação do desempenho entre BERTimbau e mBERT.

| | Fake.br | Fakepedia | FakeRecogna | FakeTrueBR |
|---|---|---|---|---|
| Fake.br | 1.0 | 0.00 | 0.0 | 0.00 |
| Fakepedia | 0.0 | 1.00 | 0.0 | 0.10 |
| FakeRecogna | 0.0 | 0.00 | 1.0 | 0.00 |
| FakeTrueBR | 0.0 | 0.22 | 0.0 | 1.00 |

Tabela 5: Resultados da sobreposição de documentos entre bases.

Após a análise, verificou-se que os dados empregados validam a hipótese de que a variação de domínio das notícias influencia negativamente a detecção de notícias falsas, isto é, contribui para o aumento da degradação dos modelos de linguagem. O desacordo entre os resultados intra-base e entre-bases evidencia a tendência dos modelos em incorporar vieses presentes nas bases, como características de vocabulário ou tendências temporais (por exemplo: dados de eleições ou da COVID-19). O baixo desempenho de generalização do *corpus FakeRecogna* evidencia que o alinhamento dos dados é um fator com maior relevância do que a dimensão da base de dados.

## 5 Conclusão

A geração massiva de informações por meio de canais digitais contribui para o compartilhamento de conteúdo inverídico, manipulado ou criado com a intenção de enganar. Surge então a necessidade de desenvolvimento de soluções capazes de tratar esse problema, assegurando às pessoas o consumo de conteúdo noticioso fidedigno. A Computação contribui nesse combate por meio do Processamento de Linguagem Natural que, a partir de modelos de linguagem, possibilita a criação de métodos eficientes para trabalhar com padrões da linguagem humana, identificando com robustez conteúdo enganoso.

O presente trabalho avaliou o fenômeno de degradação de habilidade de generalização de dois modelos baseados na arquitetura *Transformers*: BERTimbau, uma derivação do BERT pré-treinada para o português brasileiro; e o mBERT, uma versão multilíngue do BERT. Os modelos foram avaliados em quatro *corpora* de notícias falsas em português. Foi constatado que os modelos, quando ajustados em conjuntos de dados individualizados, apresentam queda na capacidade de classificar novas notícias. Isso é causado, entre outros aspectos, pela variabilidade temática dos dados, alinhamento de notícias falsas e verdadeiras deficiente ou vieses presentes nos registros que são herdados das fontes de coleta.

Por fim, a realização desta pesquisa permitiu identificar novos horizontes na análise do problema de identificação de notícias falsas em português brasileiro. Entre eles:

- conduzir a análise dos conjuntos de dados aplicando um procedimento de similaridade entre vetores de palavras, para então avaliar com maior precisão a sobreposição de informação contida nos dados, comparando com os resultados obtidos por meio da sobreposição de documentos;
- investigar o desempenho da identificação a partir do uso de arquiteturas de agentes inteligentes. Trabalhos como os de Li et al. (2024, 2025); Cui et al. (2025) e Avram et al. (2025), apresentam abordagens para tratar o problema da degradação de desempenho entre diferentes domínios e a dificuldade de modelos pré-treinados em capturar o contexto da realidade em notícias recentes;
- uma nova abordagem de treinamento de modelos, similar ao apresentado em de M. Barros et al. (2021), que simula uma segmentação temporal baseada em *chunks*, em que um texto integral verídico é segmentado em $T$ versões para serem contrastadas com as informações falsas, amplificando a capacidade preditiva do modelo. Como visto em Garcia et al. (2022); Chavarro et al. (2023) e Paixão et al. (2020), notícias verdadeiras são maiores que as falsas, podendo causar enviesamento nos resultados.

## Limitações

Este trabalho, no entanto, apresenta algumas limitações: somente uma forma de segmentação de dados foi aplicada no ajuste fino dos modelos; e os dados de notícias reais não passaram por um processo de truncamento, ou redução, para serem comparados aos inverídicos. Uma limitação intrínseca da abordagem de treinamento de modelos em bases estáticas de notícias é o fato de que esse problema de classificação é temporal em sua natureza. Ao restringir o treinamento a conjuntos de dados previamente definidos, os modelos deixam de ter acesso a informações atualizadas, reduzindo sua capacidade preditiva.

## Agradecimentos

Os autores gostariam de agradecer às agências de pesquisa brasileiras: Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG), Centro de Excelência em Inteligência Artificial (CEIA), Instituto de Informática e Universidade Federal de Goiás.

Deseja-se também externar agradecimentos aos docentes Eliomar Araújo de Lima e Jacson Rodrigues Barbosa (responsáveis pela coordenação e supervisão técnica do projeto de PD&I no qual a pesquisa está inserida); e Nádia Félix Felipe da Silva (que proveu apoio na fundamentação teórica e dos experimentos deste trabalho).

## Referências

Shadi A. Aljawarneh e Safa Ahmad Swedat. 2022. Fake news detection using enhanced BERT. *IEEE Transactions on Computational Social Systems*, 11(4):4843–4850. Publisher: IEEE.

Alexandru-Andrei Avram, Adrian Groza, e Alexandru Lecu. 2025. MCP-Orchestrated Multi-Agent System for Automated Disinformation Detection. *arXiv preprint*. ArXiv:2508.10143 [cs].

Juliana Karla de C. M. Baracho, Lucas A. Lisboa, e Roberta Vilhena V. Lopes. 2025. Levantamento e Análise Qualitativa de Bases de Dados de Fake News em Português. Em *Workshop sobre as Implicações da Computação na Sociedade (WICS)*, páginas 169–180. SBC. ISSN: 2763-8707.

Lucas Cabral, José Maria Monteiro, José Wellington Franco da Silva, César Lincoln C. Mattos, e Pedro Jorge Chaves Mourao. 2021. FakeWhastApp. BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages. Em *ICEIS (1)*, páginas 63–74.

Anderson Cordeiro Charles, Livia Ruback, e Jonice Oliveira. 2022. Fakepedia Corpus: A Flexible Fake News Corpus in Portuguese. Em *Computational Processing of the Portuguese Language*, páginas 37–45, Cham. Springer International Publishing.

Juan Pablo Chavarro, Jonata Tyska Carvalho, Tarlis Tortelli Portela, e Jonathan Cardoso Silva. 2023. FakeTrueBR: Um corpus brasileiro de notícias falsas. Em *Escola Regional de Banco de Dados (ERBD)*, páginas 108–117. SBC. ISSN: 2595-413X.

Zikun Cui, Tianyi Huang, Chia-En Chiang, e Cuiqianhe Du. 2025. Toward Verifiable Misinformation Detection: A Multi-Tool LLM Agent Framework. Em *Proceedings of the 2025 International Conference on Generative Artificial Intelligence for Business*, páginas 179–185, Hong Kong China. ACM.

Bruno de M. Barros, Hugo A. D. do Nascimento, Raphael Guedes, e Sandro E. Monsueto. 2021. Evaluating splitting approaches in the context of student dropout prediction. Em *The 17th International Conference on Frontiers in Education: Computer Science*

*& Computer Engineering (FECS'21) - to appear*, Las Vegas, USA.

Janaína Ignacio de Morais, Hugo Queiroz Abonizio, Gabriel Marques Tavares, André Azevedo da Fonseca, e Sylvio Barbon Jr. 2020. A Multi-label Classification System to Distinguish among Fake, Satirical, Objective and Legitimate News in Brazilian Portuguese. *iSys-Brazilian Journal of Information Systems (servidor antigo)*, 13(4):126–149.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, e Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. ArXiv:1810.04805 [cs].

Antônio Diogo Forte Martins, Lucas Cabral, Pedro Jorge Chaves Mourão, José Maria Monteiro, e Javam Machado. 2021. Detection of Misinformation About COVID-19 in Brazilian Portuguese WhatsApp Messages. Em Elisabeth Métais, Farid Meziane, Helmut Horacek, e Epaminondas Kapetanios, editores, *Natural Language Processing and Information Systems*, volume 12801, páginas 199–206. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Gabriel L. Garcia, Luis C. S. Afonso, e João P. Papa. 2022. FakeRecogna: A New Brazilian Corpus for Fake News Detection. Em *Computational Processing of the Portuguese Language*, páginas 57–67, Cham. Springer International Publishing.

Gabriel Lino Garcia, Pedro Henrique Paiola, Danilo Samuel Jodas, Luis Afonso Sugi, e João Paulo Papa. 2024. Text summarization and temporal learning models applied to portuguese fake news detection in a novel Brazilian corpus dataset. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, páginas 86–96.

Axel Gelfert. 2018. Fake news: A definition. *Informal Logic*, 38(1):84–117.

Hui Li, Ante Wang, Kunquan Li, Zhihao Wang, Liang Zhang, Delai Qiu, Qingsong Liu, e Jinsong Su. 2025. A Multi-Agent Framework with Automated Decision Rule Optimization for Cross-Domain Misinformation Detection. Em *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, páginas 5720–5736.

Xinyi Li, Yongfeng Zhang, e Edward C. Malthouse. 2024. Large Language Model Agent for Fake News Detection. *arXiv preprint*. ArXiv:2405.01593 [cs].

We Are Social & Meltwater. 2025. Digital 2025: Brazil.

Tomas Mikolov, Kai Chen, Greg Corrado, e Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*. ArXiv:1301.3781 [cs].

Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, e Oto A. Vale. 2018. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. Em *Computational Processing of the Portuguese Language*, páginas 324–334, Cham. Springer International Publishing.

Guilherme Zanini Moreira, Marcelo Romero, e Manassés Ribeiro. 2021. Fake News Detection Using Recurrent Neural Networks and Distributed Representations for the Portuguese Language. *sea*, 4(10):100.

Lara Souto Moreira, Gabriel Machado Lunardi, Matheus de Oliveira Ribeiro, Williamson Silva, e Fabio Paulo Basso. 2023. A study of algorithm-based detection of fake news in brazilian election: Is bert the best. *IEEE Latin America Transactions*, 21(8):897–903. Publisher: IEEE.

João Moreno e Graça Bressan. 2019. Factck.br: a new dataset to study fake news. Em *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, WebMedia '19, página 525–527, New York, NY, USA. Association for Computing Machinery.

Maik Paixão, Rinaldo Lima, e Bernard Espinasse. 2020. Fake news classification and topic modeling in Brazilian Portuguese. Em *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, páginas 427–432. IEEE.

Vinícius Baião Pires e Daniel Guerreiro. 2024. Portuguese Fake News Classification with BERT models. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, páginas 834–845. SBC.

Nils Reimers e Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Victoria L. Rubin, Yimin Chen, e Nadia K. Conroy. 2015. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Renato Moraes Silva, Hazem Amamou, Lucca Baptista Silva Ferraz, Fabio Kauê Araujo da Silva, e Anderson Raymundo Avila. 2025. Fake news detection in portuguese under large language model-generated content. *Journal of the Brazilian Computer Society*, 31(1):1150–1167.

Fábio Souza, Rodrigo Nogueira, e Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. Em *Intelligent Systems*, páginas 403–417, Cham. Springer International Publishing.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, e Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. Em *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Claire Wardle e Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making.