# Portuguese Fake News Classification with BERT models

**Vinícius Baião Pires**[1], **Daniel Guerreiro e Silva**[1]

[1]Faculdade de Tecnologia – Universidade de Brasília (UnB) – Brasília, Brasil

viniciusbaiao14@gmail.com, danielgs@unb.br

***Abstract.*** *Fake news is a phenomenum which causes great concern, due to the potential negative effects they cause in our society. Automatic fake news classification is a problem that has been addressed within machine learning community for some time, but there are not so many works on this subject which consider Portuguese as the primal language. Simultaneously, we witness the use of Transformer models with textual data, and in this context, there is BERT (Bidirectional Encoders Representations for Transformers) and its variant specifically pre-trained for portuguese tasks: BERTimbau. This work, therefore, proposes to train, with different datasets, the two aforementioned models and the multilingual variant of BERT, mBERT, for the task of classifying fake news, in order to assess the potential gains of using a language model specifically trained for Portuguese. The overall results indicate a superiority of BERTimbau over BERT and mBERT in the referred task, with an average improvement of 2.37% and 1.07%, respectively, for the F1-score.*

***Resumo.*** *O fenômeno das notícias falsas é motivo notório de preocupação, devido aos potenciais malefícios que estas causam na vida em sociedade. A classificação automática de notícias falsas (fake news) é um problema que vem sendo abordado no contexto do aprendizado de máquina há certo tempo, mas, mesmo assim, ainda não há tantos trabalhos a respeito realizados na língua portuguesa. Em paralelo, desenvolveu-se a partir dos anos 2020 o emprego de modelos Transformer com dados textuais e, neste contexto, há o modelo BERT (Bidirectional Encoders Representations for Transformers) e a sua variante que foi especificamente pré-treinada para tarefas em língua portuguesa: o BERTimbau. Este trabalho, daí, propõe treinar os dois modelos supracitados, juntos à variante multilingual do BERT, mBERT, na tarefa de classificação de notícias falsas, a partir de diversos conjuntos de dados propostos na literatura, com o objetivo de verificar os potenciais ganhos do emprego de um modelo de linguagem especificamente treinado para o Português. Os resultados obtidos indicam uma superioridade do BERTimbau sobre o BERT e sobre o mBERT na referida tarefa, com uma melhoria média de 2,37% e 1,07%, respectivamente, para os valores de F1-score.*

## 1. Introdução

Na era da informação digital, onde a disseminação de notícias ocorre instantaneamente e alcança um público global em questão de segundos, o fenômeno das notícias falsas (*fake news*) emerge como um desafio significativo. Elas são notícias fabricadas ou distorcidas que são apresentadas como fatos verídicos, com o potencial de influenciar opiniões,

moldar narrativas e até mesmo impactar decisões importantes em diversas esferas da sociedade.

Este fenômeno, porém, não é novidade. O historiador Jacob Soll indica que a origem das notícias falsas está diretamente ligada à invenção da imprensa em 1439, pois, além da aceleração da disseminação de conhecimento e informações legítimas por meio dos primeiros livros e jornais, permitiu-se também a divulgação de histórias espetaculares, como a de monstros marinhos e bruxas ou assunções de que pecadores eram responsáveis por desastres naturais [Kalsnes 2018]. Visto isso, nota-se que, ao longo da história, o fenômeno das notícias falsas foi e continua sendo comumente utilizado para benefício de agentes mal-intencionados. Isso pode ser exemplificado em períodos de eleições, em que a utilização desse recurso visa engajar eleitores mais facilmente manipuláveis [Chaves and Braga 2019].

Por isso, é de suma importância que haja a prevenção e o combate às notícias falsas. Conforme [Burkhardt 2017], algumas estratégias nesse sentido envolvem (*i*) sempre procurar saber se as pessoas que compartilham notícias com você são confiáveis ou se só estão tentando mudar sua visão sobre algo, (*ii*) sempre checar a fonte antes de compartilhar algo usando ferramentas ou *sites* de verificação, e (*iii*) sempre desconfiar de qualquer publicação. Idealmente, a adoção de tais práticas seriam suficientes para eliminar ou reduzir consideravelmente este fenômeno, o que no entanto não se verifica na sociedade como um todo. Portanto, outras estratégias de combate à divulgação de notícias falsas podem ser desenhadas, em especial, aquelas que explorem a grande disponibilidade de dados dos dias atuais e a ascensão das modernas técnicas de aprendizado de máquina. Para isso, este trabalho propõe, utilizando um modelo baseado na arquitetura *Transformer* [Vaswani et al. 2017], treinar (fazer o refinamento/ajuste-fino) e comparar três variantes (pré-treinadas com dados distintos) na tarefa de classificar um texto de notícia como sendo falso ou verdadeiro. Os modelos em questão são o BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al. 2019], junto à sua versão para múltiplas línguas mBERT [Devlin et al. 2019], e o BERTimbau [Souza et al. 2020], o qual foi pré-treinado em um *corpus* de língua portuguesa. Com o propósito de analisar o desempenho dos modelos da forma mais ampla possível, estes serão treinados e testados em vários conjuntos de dados de notícias falsas em português: Fake.Br [Monteiro et al. 2018], FakeTrueBR [Chavarro et al. 2023], FakeRecogna [Garcia et al. 2022] e Fakepedia [Charles et al. 2022].

Vale ressaltar que é a primeira vez que se aplica o BERT pré-treinado em língua portuguesa na tarefa de classificação de notícias falsas em múltiplas bases de dados, diferentemente de trabalhos anteriores, os quais empregaram o modelo BERT originalmente treinado em inglês ou a sua variante multilíngue (mBERT). Busca-se, desta forma, responder à seguinte pergunta de pesquisa: "Um modelo BERT treinado em língua portuguesa é mais eficaz na classificação de notícias falsas em português do que o BERT e o mBERT?".

O restante deste trabalho se organiza da seguinte forma: a Seção 2 aborda os trabalhos relacionados à classificação de notícias falsas, a Seção 3 apresenta o modelo BERT e a sua variante BERTimbau, a Seção 4 mostra como se deram os experimentos, mostrando os métodos e parâmetros utilizados, a Seção 5 enuncia os resultados obtidos nos experimentos e aponta as devidas observações, e, por fim, a Seção 6 apresenta as conclusões em relação a todo o trabalho.

## 2. Classificação de Notícias Falsas em Língua Portuguesa

O treinamento supervisionado de um classificador de textos na língua portuguesa pressupõe, naturalmente, a escolha de um conjunto de dados rotulado. No caso deste trabalho, selecionou-se os conjuntos Fake.Br [Monteiro et al. 2018], FakeTrueBR [Chavarro et al. 2023], FakeRecogna [Garcia et al. 2022] e Fakepedia [Charles et al. 2022]. Dentre estes, vale destacar o conjunto Fake.Br devido ao pioneirismo e à qualidade do seu processo de construção, pois, conforme [Monteiro et al. 2018] detalham, as 7200 notícias que o compõem são alinhadas, i.e. para cada notícia verdadeira há uma falsa que a contrapõe, tornando o conjunto mais adequado para que o modelo aprenda a discriminar o conteúdo legítimo de sua versão falsa. Além disso, o processo de seleção das notícias falsas foi exclusivamente feito de forma manual, com a devida verificação, enquanto que, para as notícias verdadeiras, foram pré-coletadas 40.000 notícias de algumas das grandes agências de notícias do Brasil e, após essa coleta, foram escolhidas de forma manual aquelas que se contrapunham à notícia falsa previamente escolhida.

Desde a criação do Fake.Br, o mesmo tem sido utilizado como referência para validar diversos modelos de classificação de notícias falsas. Em [Silva et al. 2020], foram aplicados alguns métodos de classificação como regressão logística, máquina de vetores-suporte, árvores de decisão, floresta aleatória, entre outros. Com isso, nos experimentos utilizando um vetor de atributos linguísticos da notícia como entrada, os modelos obtiveram *F1-score* maior que $0{,}9$, enquanto que nos experimentos feitos tomando como entrada a representação *Bag of Words* (BoW) do texto inteiro, obteve-se um *F1-score* de $0{,}937$. Por fim, ao associar as duas estratégias, atingiu-se um resultado de $0{,}965$, o que representou um ganho de, aproximadamente, $9{,}7\%$ em relação ao melhor resultado obtido em [Monteiro et al. 2018].

Recentemente, [Fischer et al. 2022] propuseram classificar notícias falsas em português usando, pela primeira vez, um modelo *Transformer*, especificamente o mBERT. O conjunto Fake.Br foi objeto dos experimentos, e também se fez uso de métodos de classificação semelhantes aos vistos em [Silva et al. 2020], como a representação BoW e separação de cenários com os textos nas formas truncada ou na sua forma completa. O melhor desempenho obtido correspondeu a um *F1-score* de $98{,}4\%$, indicando a possível superioridade do uso de *Transformers* em relação a outros métodos nesta tarefa.

Apesar do fenômeno das notícias falsas continuar a ser bastante repercutido no mundo todo, a maioria dos trabalhos relacionados à sua classificação utilizando aprendizado de máquina são direcionados a corpus em inglês e, por isso, criar outros corpus confiáveis como o Fake.Br continua necessário. Desta forma, a maioria dos trabalhos na área voltados para a língua portuguesa envolvem criar modelos melhores na classificação de notícias falsas com o Fake.Br, como visto em [Silva et al. 2020] e [Fischer et al. 2022], ou a criação de novos conjuntos de dados, como os que vem a seguir.

**Fakepedia** [Charles et al. 2022]: neste corpus, além das notícias falsas, foi coletado e processado conteúdo de notícias reais, extraídos de portais confiáveis da internet. Mas a contribuição principal do trabalho é a apresentação de um processo automático de criação da base, o que permite que o mesmo se mantenha relativamente atualizado. Apesar de ser uma boa alternativa, o processo de seleção e verificação automática de notícias falsas/verdadeiras leva a uma menor confiabilidade na qualidade do conjunto como um

todo, em comparação ao processo manual do Fake.Br. Além disso, notícias verdadeiras que fazem parte do conjunto que deveria contrapor uma notícia falsa podem estar falsamente relacionadas, pois a estratégia de associação entre elas pode envolver uma única palavra, o que não seria suficiente para se comparar contextos e se garantir uma relação do ponto de vista semântico.

**FakeRecogna** [Garcia et al. 2022]: assim como o Fakepedia, este conjunto busca trazer notícias mais recentes que as presentes no Fake.Br, coletadas automaticamente entre os anos de 2019 e 2021. Diferentemente do Fake.Br, este conjunto não apresenta a característica de que, para cada notícia falsa, há uma verdadeira que a contrapõe, o que faz com que o treinamento de um modelo neste conjunto possa falhar no aprendizado dos padrões e regularidades eventualmente existentes quando se comparam exemplos "positivos" e "negativos".

**FakeTrueBR** [Chavarro et al. 2023]: para a criação deste conjunto foi utilizado o site *Boatos.org*[1], que captura mensagens falsas e, além disso, reporta o motivo pelo qual a notícia é falsa, mas não cita uma fonte que justificaria a sua falsidade. Dessa forma, utilizando um *crawler*, foram selecionadas 4600 notícias falsas, das quais 2083 continham mais de 300 caracteres, um requisito para cada notícia falsa presente no conjunto. Para a seleção das notícias verdadeiras, foi novamente utilizado um *crawler* para vasculhar os sites confiáveis de notícias G1 e Folha, resultando na recuperação de 3032 notícias verdadeiras, de forma que cada uma destas contraponha alguma notícia falsa. Por fim, o conjunto ficou com 3582 notícias, igualmente divididas entre verdadeiras e falsas. Apesar deste conjunto possuir notícias mais recentes que o Fake.Br, entre os anos de 2017 a 2023, suas notícias não foram selecionadas manualmente e, além disso, apesar das notícias falsas possuírem obrigatoriamente mais de 300 caracteres, as verdadeiras não necessariamente cumprem essa condição, contribuindo, assim, para uma possível maior dificuldade para um modelo, após treinado nesta base, conseguir classificar notícias verdadeiras.

Dado o contexto dos trabalhos supracitados, nota-se que diversos modelos já foram treinados e validados com o corpus Fake.Br (e com outros corpus), inclusive, os modernos modelos baseados na arquitetura *Transformer*, como o BERT e o mBERT. No entanto, nota-se a ausência de um estudo do desempenho do modelo BERTimbau, que se distingue por ser uma variante do BERT pré-treinada em um corpus de textos em português. Então, como contribuição, este trabalho propõe treinar (realizar o ajuste-fino) o modelo de linguagem BERTimbau nos quatro conjuntos de notícias falsas/verdadeiras já apresentados (Fake.Br, Fakepedia, FakeRecogna e FakeTrueBR) e comparar os seus desempenhos com os equivalentes aos modelos BERT e mBERT, treinados sob as mesmas condições, com a intenção de elucidar se o uso de modelos de linguagem especificamente pré-treinados em língua portuguesa contribui para a melhoria do desempenho na classificação de notícias falsas, na referida língua. Ademais, testaremos os modelos com tal conjunto diversificado de bases na intenção de confirmar a robustez dos modelos baseados na arquitetura *Transformer* para a tarefa em foco, neste artigo.

## 3. Os modelos BERT, mBERT e BERTimbau

Baseados na arquitetura Transformer [Vaswani et al. 2017], pesquisadores propuseram em 2018 o BERT [Devlin et al. 2019], um modelo grande de linguagem para inúmeras

[1] `https://www.boatos.org/`

aplicações de processamento de linguagem natural, entre elas a classificação de textos.

Sua arquitetura se trata de um codificador *Transformer* bidirecional multicamadas, o que, de forma resumida, permite ao modelo receber como entrada textos completos de uma só vez, ao invés da forma sequencial requerida pelas redes neurais recorrentes, ao mesmo tempo que explora com competência os mecanismos de atenção e representação dos *tokens* em espaço latente já advindos de estratégias anteriores. O modelo base é composto por uma pilha de 12 codificadores *Transformer*. Cada bloco deste possui internamente duas subcamadas, com a primeira sendo um *multi-head self-attention mechanism* e a segunda sendo uma rede *feedforward* simples totalmente conectada. É então empregada uma conexão residual em torno de cada uma das duas subcamadas, seguida por uma técnica de *Layer Normalization*.

A aplicação de um modelo de linguagem como o BERT em um tarefa usualmente considera dois passos: o pré-treino e o refinamento/ajuste-fino (*fine-tuning*). No pré-treino, o modelo passa por um treinamento não-supervisionado em duas tarefas genéricas. Já no ajuste-fino, o modelo é iniciado com os parâmetros pré-treinados, os quais continuam a ser ajustados usando dados rotulados da tarefa-fim (*downstream task*) [Saunshi et al. 2021], como classificação de textos. A Figura 1 apresenta a diferenciação entre as etapas de pré-treino e ajuste-fino do BERT. O BERT está disponível já pré-treinado em dois tamanhos, sendo eles o BERTBASE, com 110 milhões de parâmetros, e o BERTLARGE, com 340 milhões. Seu sucesso pode em parte se explicar pelo emprego no pré-treino do aprendizado da tarefa de *Masked Language Model* (MLM) [Devlin et al. 2019], na qual aleatoriamente se mascaram *tokens* da entrada e se faz o modelo predizer qual seria o *token* original baseado nos elementos ao seu redor, o que é favorecido pela sua capacidade de processar a sequência lida por inteiro nos dois sentidos, de forma que o modelo possa adquirir contexto tanto pelo fim como pelo começo da entrada.

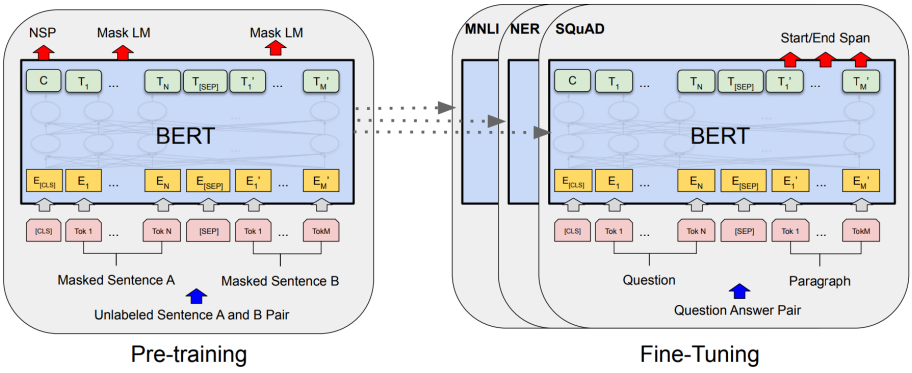


**Figura 1. Procedimentos gerais de pré-treino e ajuste-fino para o BERT. Retirado de [Devlin et al. 2019].**

A variante multilíngue mBERT [Devlin et al. 2019], por outro lado, é um modelo que foi pré-treinado num corpus formado pelo conteúdo das *Wikipedias* nas 104 línguas com maiores quantidades de verbetes. No momento, este modelo possui duas formas: *Uncased* e *Cased*, em que a segunda é mais recomendada para uso por ser a versão mais recente e, com isso, possuir diversas otimizações, como a correção de alguns problemas de normalização que existiam em algumas línguas não-latinas. Em termos de dimensão,

apenas suas respectivas versões BASE estão disponíveis para uso.

Em 2020, pesquisadores brasileiros realizaram o pré-treino de modelos BERT para a língua portuguesa, chamando tais variantes de BERTimbau [Souza et al. 2020]. Para construí-las, utilizou-se do corpus brWaC [Wagner Filho et al. 2018], até então o maior corpus aberto de textos em português. O BERTimbau também é disponibilizado em dois tamanhos, BASE e LARGE, equivalentes às versões de mesmo nome do BERT, e pré-treinados por meio das mesmas tarefas.

A fim de ilustrar a sua eficácia, apresentamos aqui a avaliação do modelo em 3 tarefas-fim: *Sentence Textual Similarity* (STS), *Recognizing Textual Entailment* (RTE) e *Named Entity Recognition* (NER). Os resultados estão nas Tabelas 1 e 2, os quais indicam que o BERTimbau consegue superar resultados estabelecidos anteriormente por modelos multilíngue equivalentes nas tarefas propostas. Por isso, e com base na motivação previamente explicitada ao fim da Seção 2, neste trabalho será aplicado, pela primeira vez, o BERTimbau na tarefa de classificação de notícias falsas de modo a se comparar os seus resultados com aqueles atingidos a partir da classificação com o BERT e o mBERT.

**Tabela 1. Resultados do BERTimbau nas tarefas STS e RTE [Souza et al. 2020].**

| **Modelo** | **STS** | | **RTE** | |
|---|---|---|---|---|
| | ***Correlação de Pearson*** | ***MSE*** | ***F1-Score*** | ***Acurácia*** |
| BERTimbau BASE | 0,836 | 0,58 | 89,2 | 89,2 |
| BERTimbau LARGE | 0,852 | 0,50 | 90,0 | 90,0 |

**Tabela 2. Resultados do BERTimbau na tarefa NER [Souza et al. 2020].**

| **Modelo** | **NER** | | | | | |
|---|---|---|---|---|---|---|
| | **Cenário Total** | | | **Cenário Seletivo** | | |
| | ***Prec.***[a] | ***Rec.***[b] | ***F1-Score*** | ***Prec.***[a] | ***Rec.***[b] | ***F1-Score*** |
| BERTimbau BASE | 76.8 | 77.1 | 77.2 | 81.9 | 82.7 | 82.2 |
| BERTimbau LARGE | 77.9 | 78.0 | 77.9 | 81.3 | 82.2 | 81.7 |

[a]Precisão [b]Recall

## 4. Experimentos

Nesta seção mostramos os procedimentos usados para realização dos experimentos, os quais possuem como objetivo a comparação dos modelos BERTimbau, BERT e mBERT na tarefa de classificação de notícias falsas em português.

Estes modelos são treinados nos corpus Fake.Br, Fakepedia, FakeRecogna e FakeTrueBR utilizando as respectivas versões BASE. Isto se dá devido a limitações de custo computacional e para fins de justiça, uma vez que não há versão LARGE do modelo mBERT. A implementação[2] é feita na linguagem Python [Van Rossum 1991]. O experimento é executado num computador com as seguintes características: processador AMD

[2]https://github.com/Baiaopires/Portuguese-Fake-News-Classification-with-BERT-models

Ryzen Threadripper 3990X de 64 cores, 152GB de memória RAM e GPU NVIDIA GeForce RTX 3090 com 24GB de memória GDDR6X. A cada rodada de aferição dos resultados, os corpus são particionados nos conjuntos de treino, validação e teste, enquanto que a métrica de comparação entre os modelos envolve a acurácia, o *recall* e *F1-score* calculados nos conjuntos de teste.

### 4.1. Pré-processamento

Primeiramente, é necessário realizar algumas rotinas de pré-processamento dos textos de cada corpus. Estas ações são idênticas para os três modelos em consideração, uma vez que estes possuem a mesma arquitetura, só se distinguindo pela forma em que se deu o pré-treino de cada um.

O conjunto Fake.Br é obtido originalmente no formato CSV (*Comma Separated Values*), contendo três campos: *index*, *label* e *preprocessed news*. Necessita-se, então, apenas transferir os dados do conjunto para uma matriz bidimensional, de forma que as colunas fiquem separadas em notícias (*preprocessed news*) e rótulos (*label*), sendo utilizado, para este fim, a biblioteca Pandas [McKinney et al. 2011]. Como é possível notar pelo seu nome, o campo *preprocessed news* já traz os dados com algum tipo de pré-processamento. Neste caso, as notícias já passaram por uma etapa de normalização, que consistiu na conversão em letras minúsculas e na remoção de *stopwords* e de acentuação. Ainda que tal remoção seja questionável para modelos de linguagem que trabalham com contexto, como o BERT, os dados na base já são apresentados com esse processamento porque a sua ausência pode prejudicar o desempenho principalmente de modelos de aprendizado de máquina "tradicionais" [Clark and Araki 2011]. O conjunto Fakepedia, diferentemente do Fake.Br, não é, a princípio, totalmente preenchido, de forma que alguns exemplos, apesar de existirem, não possuem uma das partes fundamentais para os experimentos, que é o texto das notícias. Estes exemplos, naturalmente, foram eliminados da matriz de dados e, dessa forma, o tamanho efetivo do conjunto passou de 13.822 para 12.326 notícias. Para o conjunto FakeRecogna, detectou-se que apenas um exemplo da base original tinha o texto da notícia ausente, o qual foi descartado e, com isso, o conjunto pôde ser utilizado após manipulações similares às feitas no conjunto Fake.Br. Por último, as notícias do conjunto FakeTrueBR estavam distribuídas de forma que ambas as notícias falsa e sua contraparte verdadeira compunham a mesma linha da matriz, sem possuir coluna específica de rótulo (*label*), houve então a necessidade da inserção desta informação previamente ao uso.

Por fim, há o processo de *tokenization* [Tunstall et al. 2022] e codificação vetorial dos textos, necessária para transformar cada documento na forma de entrada que é aceita pelos modelos. Cada uma das palavras presentes na notícia é separada de forma que todas estejam em entradas diferentes na matriz. Em seguida, cada palavra é substituída por um vetor (*embedding*) que codifica o seu significado mas também a sua posição no texto (*positional encoding*) [Vaswani et al. 2017]. Foi utilizado o tamanho máximo de 128 tokens para cada notícia (linha da matriz tokenizada).

### 4.2. Treinamento (Ajuste-fino)

Antes de treinar os modelos, é necessário separar os conjuntos de treino, validação e teste. Como já mencionado, a distribuição entre as duas classes de notícias (falsa/verdadeira) é uniforme, daí se realizou a separação entre os três conjuntos de forma estratificada, i.e.

para que se mantivesse a mesma distribuição de classes. Por fim, a proporção escolhida de amostras para os conjuntos foi de 60%, 20%, e 20% do total para treino, validação e teste, respectivamente.

Utilizou-se o otimizador AdamW [You et al. 2020] com *learning rate* de $5 \times 10^{-2}$ e $\epsilon = 5 \times 10^{-8}$. O treino então se deu de maneira que os modelos eram inicializados com os seus parâmetros pré-treinados, os quais continuavam a ser ajustados através da minimização da função custo de entropia cruzada sobre o conjunto de treino, com um tamanho de *batch* igual a 128 e número de épocas determinado automaticamente por meio da técnica de *early-stopping* [Prechelt 1998] junto ao conjunto de validação, com paciência igual a 100 épocas. Este valor de paciência foi definido em um experimento preliminar, que será descrito na Seção 5, e tem o objetivo de permitir a cada modelo um tempo de treinamento suficientemente grande e sem risco de detectar a estagnação do treinamento (e portanto a sua parada) devido a oscilações momentâneas na medida da métrica, com a ressalva de que não se corre o risco de sobreajuste porque o modelo ao final do treinamento é recuperado da época em que houve o melhor desempenho junto ao conjunto de validação.

### 4.3. Testes

Encerrado o processo de ajuste-fino, cada modelo é avaliado junto ao conjunto de testes para verificar se atingiu o desempenho adequado, em termos de generalização, na classificação de notícias falsas. Como já mencionado, as métricas de qualidade consideradas são a acurácia, o *recall* e o *F1-score* nos referidos corpus.

## 5. Resultados

Iniciamos com um experimento preliminar a fim de definir o valor do hiperparâmetro de paciência do método de *early-stopping*, uma vez que é esperado haver uma variabilidade na duração, em termos de número de épocas, de cada rodada de treinamento, para cada modelo, devido à diferença no conjunto inicial de parâmetros de cada arquitetura e devido ao particionamento dos conjuntos de treino, validação e teste ser feito de forma aleatória.

Então, foi feito um treinamento do modelo BERTimbau com um valor elevado de paciência, no caso 1000 épocas, para se analisar de forma qualitativa o comportamento da curva de acurácia junto ao conjunto de validação. Como pode ser visto pela Figura 2, após um valor próximo de 400 épocas, a acurácia cai de uma média de $0{,}95$ para cerca de $0{,}93$ e, ao redor da época 950, o mesmo fenômeno ocorre, chegando a uma nova medida de $0{,}87$. Finalmente, o algoritmo decide parar o treinamento na época 1027. Lembrando que o valor configurado para o *early stopping* neste experimento preliminar havia sido de 1000, isto significa que o modelo não viu nenhuma evolução significativa na sua acurácia desde a época 27. Uma vez que não há garantias de que outras execuções teriam uma duração reduzida de épocas, ao mesmo tempo que este resultado indica que não haveria necessidade de grandes períodos de treinamento, decidiu-se como valor de compromisso a ser utilizado nos experimentos subsequentes 100 épocas de paciência.

### 5.1. Experimento principal

Passou-se, finalmente, aos experimentos principais de comparação entre os modelos BERT, mBERT e BERTimbau. Foram executados 10 experimentos independentes de

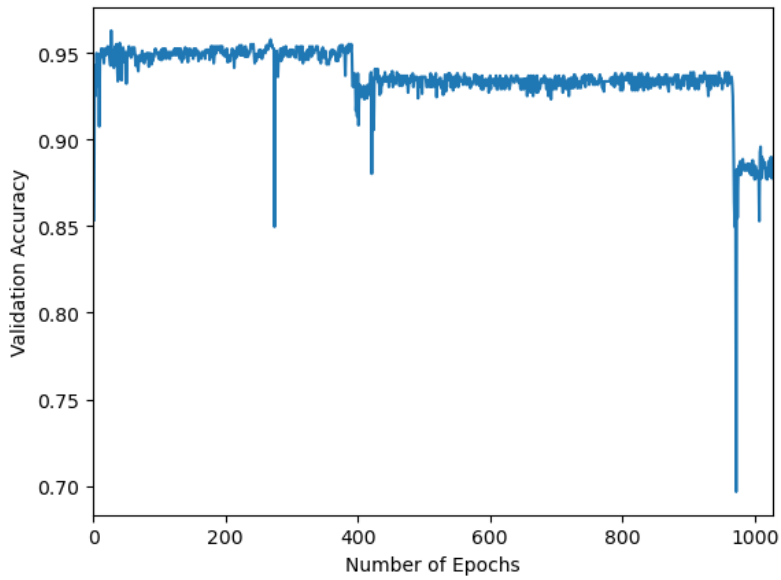


**Figura 2. Acurácia de validação em um experimento preliminar com o modelo BERTimbau e o conjunto Fake.Br, com 1000 épocas de paciência.**

treinamento e teste para cada modelo e conjunto de dados. Para melhor entendimento geral dos resultados, foi calculada a média e desvio-padrão das métricas sobre todos os experimentos, como pode ser visto na Tabela 3.

Como é possível perceber pela medida de acurácia, é notável que, independentemente da base de dados, as notícias foram melhor classificadas em média pelo modelo BERTimbau. Vale lembrar que todos os experimentos foram feitos com a mesma paciência para o *early stopping*, alguns destes, inclusive, obtiveram resultados melhores mesmo consumindo menos épocas para finalizar, o que indica que o método de parada foi suficiente para garantir um treinamento adequado. Percebe-se também, pela medida de *Recall*, que o número de verdadeiros positivos é muito maior que o dos falsos negativos, o que indica que os modelos conseguem com bastante frequência classificar corretamente os casos em que as notícias são verdadeiras.

Ademais, os valores elevados de Precisão também indicam que o alto *Recall* foi atingido sem incorrer em muitos falsos positivos, i.e. notícias falsas erroneamente classificadas como legítimas. Com a Precisão e o *Recall*, foi possível calcular o *F1-score*, o qual integra os dois valores em apenas uma métrica e, com isso, traz um importante indicador de equilíbrio entre as duas informações, permitindo, assim, um melhor entendimento a respeito do desempenho dos modelos. Pela Tabela 3, observa-se que todos os modelos alcançaram um bom equilíbrio entre as duas medidas, mas que, da mesma forma que antes, houve uma superioridade do modelo brasileiro em relação aos demais.

Um ponto a se destacar destes resultados, também, é a proximidade dos valores do modelo mBERT com as do BERTimbau, lembrando que, diferentemente deste, o mBERT foi pré-treinado com o objetivo de ser útil para uma grande variedade de outras línguas. Esta proximidade, inclusive, em alguns dos casos se apresenta pela sobreposição dos intervalos de confiança. De todo modo, sob a perspectiva da média, em nenhum cenário o mBERT conseguiu superar o BERTimbau. Além disso, como também é perceptível, todos

**Tabela 3. Resultados para cada modelo e cada corpus. Média e desvio-padrão das métricas (junto ao conjunto de teste) em 10 rodadas independentes de treinamento.**

| *Dataset* | Modelo | Média ± Desvio Padrão | | | |
|---|---|---|---|---|---|
| | | *Acurácia* | *Precisão* | *Recall* | *F1-score* |
| Fake.Br | BERT | $0{,}897 \pm 0{,}005$ | $0{,}896 \pm 0{,}013$ | $0{,}896 \pm 0{,}014$ | $0{,}896 \pm 0{,}009$ |
| | mBERT | $0{,}938 \pm 0{,}005$ | $0{,}937 \pm 0{,}013$ | $0{,}939 \pm 0{,}013$ | $0{,}938 \pm 0{,}005$ |
| | BERTimbau | $\mathbf{0{,}950 \pm 0{,}002}$ | $\mathbf{0{,}955 \pm 0{,}007}$ | $\mathbf{0{,}945 \pm 0{,}009}$ | $\mathbf{0{,}950 \pm 0{,}003}$ |
| FakeTrueBR | BERT | $0{,}965 \pm 0{,}004$ | $0{,}963 \pm 0{,}011$ | $0{,}965 \pm 0{,}007$ | $0{,}964 \pm 0{,}004$ |
| | mBERT | $0{,}974 \pm 0{,}006$ | $0{,}973 \pm 0{,}015$ | $0{,}976 \pm 0{,}009$ | $0{,}975 \pm 0{,}006$ |
| | BERTimbau | $\mathbf{0{,}981 \pm 0{,}003}$ | $\mathbf{0{,}979 \pm 0{,}007}$ | $\mathbf{0{,}984 \pm 0{,}007}$ | $\mathbf{0{,}982 \pm 0{,}003}$ |
| FakeRecogna | BERT | $0{,}951 \pm 0{,}004$ | $0{,}949 \pm 0{,}009$ | $0{,}954 \pm 0{,}007$ | $0{,}951 \pm 0{,}004$ |
| | mBERT | $0{,}963 \pm 0{,}003$ | $0{,}962 \pm 0{,}008$ | $0{,}963 \pm 0{,}008$ | $0{,}963 \pm 0{,}003$ |
| | BERTimbau | $\mathbf{0{,}970 \pm 0{,}002}$ | $\mathbf{0{,}971 \pm 0{,}006}$ | $\mathbf{0{,}968 \pm 0{,}004}$ | $\mathbf{0{,}969 \pm 0{,}002}$ |
| Fakepedia | BERT | $0{,}969 \pm 0{,}003$ | $0{,}969 \pm 0{,}006$ | $0{,}969 \pm 0{,}004$ | $0{,}969 \pm 0{,}003$ |
| | mBERT | $0{,}981 \pm 0{,}003$ | $0{,}982 \pm 0{,}003$ | $0{,}980 \pm 0{,}005$ | $0{,}981 \pm 0{,}003$ |
| | BERTimbau | $\mathbf{0{,}985 \pm 0{,}002}$ | $\mathbf{0{,}985 \pm 0{,}005}$ | $\mathbf{0{,}986 \pm 0{,}004}$ | $\mathbf{0{,}985 \pm 0{,}002}$ |

os modelos apresentaram maior dificuldade para classificar o conjunto de dados Fake.Br, como pode ser notado quando se comparam os valores de *Recall* e Acurácia e, ainda neste mesmo conjunto, houve o maior salto de desempenho do BERTimbau quando comparado ao BERT. A causa disso pode estar relacionada com a maior qualidade do processo de construção deste conjunto, em termos de menos etapas automatizadas de coleta de dados e alinhamento mais confiável de notícia verdadeira *versus* falsa, quando comparada com as demais bases.

Por fim, é possível perceber que os resultados alcançados pelo BERTimbau são mais consistentes que os do BERT e que os do mBERT. Pelo valor das médias, nota-se um ganho médio de $2{,}64\%$ do BERTimbau em relação ao BERT e $1{,}17\%$ em relação ao mBERT no *Recall*, $2{,}8\%$ e $0{,}78\%$, respectivamente, na Acurácia e $2{,}37\%$ e $1{,}07\%$ para os valores do *F1-score*, o que, portanto, permite concluir que o uso de um modelo de linguagem pré-treinado em português, o BERTimbau, produziu resultados nitidamente melhores na tarefa de classificação de notícias falsas na referida língua, quando comparado ao seu modelo precursor e sua variante multilíngue, o BERT e o mBERT.

## 6. Conclusões

Neste trabalho, a partir da necessidade de expandir ainda mais o conhecimento a respeito da classificação automática de notícias falsas, pode-se atestar, por meio dos conjuntos de dados Fake.Br, Fakepedia, FaketTrueBR e FakeRecogna, que é possível realizar a referida tarefa na língua portuguesa de forma mais precisa com a utilização de um modelo *Transformer* especificamente pré-treinado para o português, BERTimbau, ao invés do modelo original pré-treinado em língua inglesa, que é o caso do BERT, ou de sua versão pré-treinada em 104 linguagens distintas, o mBERT.

Com os resultados obtidos presentes na Tabela 3, foi possível perceber um incremento médio de $2{,}8\%$ e $0{,}78\%$ na Acurácia do BERTimbau em relação ao BERT e o mBERT, respectivamente. Além disso, houve um ganho médio de $2{,}64\%$ e $1{,}17\%$ no *Recall* para o modelo brasileiro sobre o BERT e o mBERT, respectivamente, o que mostra uma proporção maior de verdadeiros positivos em relação aos falsos negativos, contribuindo, assim, para uma maior confiança com relação à classificação correta das notícias verdadeiras. É perceptível também o ganho de $2{,}37\%$ e $1{,}07\%$ para os valores do *F1-score* sobre os mesmos modelos, que mostra um maior equilíbrio entre os valores de Precisão e *Recall* em relação aos outros modelos.

Estes resultados respondem positivamente à pergunta de pesquisa apresentada no início deste trabalho. De todo modo, ainda há certamente espaço para trabalhos futuros, como uma investigação envolvendo treinamento e testes com bases distintas i.e. *out-of-distribution*, a aplicação de testes estatísticos na análise experimental, além da incorporação de atributos de entrada não textuais (metadados das notícias e/ou atributos linguísticos).

## Referências


Burkhardt, J. M. (2017). *Combating fake news in the digital age*. Number vol. 53, no. 8 in Library technology reports. ALA TechSource, Chicago, IL.

Charles, A. C., Ruback, L., and Oliveira, J. (2022). Fakepedia Corpus: A Flexible Fake News Corpus in Portuguese. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 37–45, Cham. Springer International Publishing.

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

Chaves, M. and Braga, A. (2019). The agenda of disinformation: "fake news"and membership categorization analysis in the 2018 Brazilian presidential elections. *Brazilian journalism research*, 15(3):474–495.

Clark, E. and Araki, K. (2011). Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. *Procedia - Social and Behavioral Sciences*, 27:2–11.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, volume 1, pages 4171–4186, Stroudsburg, PA, USA. Association for Computational Linguistics. arXiv: 1810.04805.

Fischer, M., Haque, R., Stynes, P., and Pathak, P. (2022). Identifying Fake News in Brazilian Portuguese. In *Natural Language Processing and Information Systems*, pages 111–118, Cham. Springer International Publishing.

Garcia, G. L., Afonso, L. C. S., and Papa, J. P. (2022). FakeRecogna: A New Brazilian Corpus for Fake News Detection. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 57–67, Cham. Springer International Publishing.

Kalsnes, B. (2018). Fake News. In *Oxford Research Encyclopedia of Communication*. Oxford University Press.

McKinney, W. et al. (2011). pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.

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 Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., and Paetzold, G. H., editors, *Computational Processing of the Portuguese Language*, pages 324–334, Cham. Springer International Publishing.

Prechelt, L. (1998). Early Stopping - But When? In Orr, G. B. and Müller, K.-R., editors, *Neural Networks: Tricks of the Trade*, pages 55–69. Springer Berlin Heidelberg, Berlin, Heidelberg.

Saunshi, N., Malladi, S., and Arora, S. (2021). A mathematical exploration of why language models help solve downstream tasks. In *International Conference on Learning Representations*.

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

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12319 LNAI:403–417. ISBN: 9783030613761.

Tunstall, L., Von Werra, L., and Wolf, T. (2022). *Natural language processing with transformers*. "O'Reilly Media, Inc.".

Van Rossum, G. (1991). Python programming language. `https://www.python.org/`. Acessado em: 23-08-2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. (2020). Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*.