Classificação de Notícias Falsas na Língua Portuguesa Utilizando Modelos Baseados na Arquitetura Transformer

Lucas G. Pellegrini[1], **Fernanda M. C. Santos**[1], **Felipe H. S. Cantarino**[1]

[1]Faculdade de Computação – Universidade Federal de Uberlândia (UFU)
Uberlândia – MG – Brasil

lucas.pellegrini@ufu.br, fmcsantos@ufu.br, harrisufelipe@gmail.com

Abstract. *The quick growth of internet and social media usage has contributed to the widespread dissemination of the so-called Fake News. The alarming proportions this phenomenon has reached suggests the existence of a gap in the fight against misinformation. This study aims to employ classification models based on the Transformer neural network architecture for the task of fake news classification in texts written in Portuguese. Therefore, three distinct models were developed: (1) Encoder-Only, (2) Decoder-Only, and (3) Transformer (Encoder-Decoder); all trained on the same dataset obtained by merging two corpora. Additionally, some pre-trained models were analyzed and their results compared with those of the proposed models. In summary, all developed Transformer models demonstrated superior performance, with particular emphasis on the Encoder-Only model, which achieved accuracy and precision values exceeding 96.7%.*

Resumo. *A rápida expansão do uso da internet e das redes sociais tem contribuído para a disseminação das chamadas "Fake News" (Notícias Falsas). As proporções que esse fenômeno tomou sugerem a existência de uma lacuna no combate à desinformação. Assim, este trabalho tem como objetivo empregar modelos de classificação baseados na arquitetura das redes neurais Transformer na tarefa de classificação de notícias falsas em textos escritos na língua Portuguesa. Para isso, três modelos distintos foram desenvolvidos: (1) Encoder-Only, (2) Decoder-Only e (3) Transformer (Encoder-Decoder); todos treinados sobre um mesmo conjunto de dados oriundo da união de dois corpora. Ademais, foram analisados alguns modelos pré-treinados e comparados seus resultados com os dos modelos propostos neste artigo. Em suma, todos os modelos Transformer desenvolvidos apresentaram desempenho superior, com destaque para o modelo Encoder-Only, que obteve valores de acurácia e precisão superiores a 96,7%.*

1. Introdução

A democratização do acesso à internet faz com que, cada vez mais, pessoas tenham a capacidade de se conectar, compartilhar e acessar informações, seja por meio de veículos de notícias, seja através de redes sociais. Esse crescimento, que é uma realidade no Brasil [IBGE 2023], é acompanhada por uma problemática: a disseminação de notícias falsas (popularmente conhecidas pelo termo em inglês *"fake news"*). Segundo [Poynter 2022], mais de quatro a cada dez brasileiros acreditam encontrar informações falsas diariamente,

o que evidencia a relevância do problema, bem como a importância de estudar e desenvolver técnicas para conter a propagação da desinformação.

Uma maneira de lidar com o problema abordado é a classificação de notícias como verdadeiras ou falsas, como faz o serviço do grupo Globo [G1 - Fato ou Fake 2025]. Na literatura, encontra diferentes estudos que descrevem o uso de técnicas de Processamento de Linguagem Natural (PLN) aliada à Redes Neurais para fazer a análise e a classificação de notícias [Narde et al. 2024, Pires and e Silva 2024].

Assim sendo, este artigo propõe desenvolver e treinar modelos baseados na arquitetura *Transformer*, originários do trabalho de [Vaswani et al. 2017], que são: (*Encoder-Decoder*), *Encoder-Only* e *Decoder-Only*. Os resultados obtidos por esses modelos serão analisados e comparados. Ademais, modelos pré-treinados na classificação de notícias falsas na língua Portuguesa também serão apresentados e confrontados com os resultados obtidos com os modelos anteriores.

2. Modelo Proposto

2.1. Conjunto de dados

Neste trabalho, foram utilizadas duas bases de dados compostas por textos em português: (1) BoatosBrCorpus[1] e (2) FakeBr[2]. A primeira base, é formada por 3427 notícias do site `Boatos.org`, sendo 1911 classificadas como falsas e 1516 verdadeiras. A segunda base de dados, por sua vez, é constituída por 7200 notícias, igualmente distribuídas entre verdadeiras e falsas, coletadas dos sites *Diário do Brasil*, *A Folha do Brasil*, *The Jornal Brasil* e *Top Five TV*.

A utilização em conjunto das duas bases de dados não só expande o tamanho do acervo como também o diversifica. Notícias provenientes de diferentes fontes, jornalísticas ou não, trazem consigo uma variedade de vocabulário e assuntos abordados. A FakeBr aborda temas como política, TV e celebridades, sociedade e notícias do dia-a-dia, ciência e tecnologia, economia e religião; enquanto a BoatosBrCorpus traz matérias sobre Brasil, mundo, ciência, entretenimento, saúde, tecnologia, esportes, economia e religião.

2.2. Pré-processamento e Embedding

Com o intuito de preparar os dados para os modelos de classificação, buscou-se reduzir eventuais discrepâncias entre os textos presentes nas diferentes bases. Assim, foram adotadas as técnicas de derivação (do inglês "*stemming*") e de lematização (do inglês "*lemmatization*"), que diminuem a carga de processamento realizada pelos modelos. A derivação consiste em reduzir uma palavra até sua forma raiz/base, enquanto a lematização é o processo de associar palavras similares a um lema (uma outra palavra) [Khyani and B S 2021].

As etapas adotadas no pré-processamento foram: (1) Remoção de acentos, *stopwords*[3], caracteres especiais e pontuação; (2) Lematização [4]; (3) Derivação, [5]; (4)

[1]Disponível em: https://github.com/Felipe-Harrison/boatos-br-corpus [Cantarino 2024]
[2]Disponível em: https://github.com/roneysco/Fake.br-Corpus [Monteiro et al. 2018]
[3]Palavras pouco significativas
[4]tarefa realizada com auxílio da biblioteca *spaCy*
[5]tarefa realizada com auxílio da biblioteca Natural Language Toolkit (NLTK)

Remoção de linhas com colunas nulas. O resultado desta etapa foi uma base de dados com 10604 notícias, das quais 5483 são classificadas como falsas e 5121 como verdadeiras.

Na sequência, as notícias pré-processadas foram transformadas em sequências numéricas para serem utilizadas como dados de entrada aos modelos de classificação baseados em redes neurais. Para isso, foi aplicado o processo simples de *embedding* do tipo *text-to-sequence*. Nesse processo, o texto de cada notícia é primeiro transformado em vetores (listas de *tokens*), utilizados para gerar um vocabulário global que, em seguida, é utilizado para mapear cada *token*. Posteriormente, o processo de *padding* garante que todas as sequências configurem, exatamente, o mesmo tamanho.

2.3. Modelo de Classificação

Neste trabalho, foram desenvolvidos três implementações baseadas na arquitetura *Transformers*, visando uma compreensão mais aprofundada de sua estrutura e funcionamento. São eles: (1) *Encoder-Only*, modelo constituído apenas da camada codificadora da arquitetura original, conforme ilustrado na Figura 1; (2) *Decoder-Only*, modelo constituído apenas da camada decodificadora da arquitetura original, conforme ilustra a Figura 2; (3) *Transformer*, modelo constituído pela união dos modelos (1) e (2), onde as duas componentes recebem os dados de entrada, mas a saída da componente codificadora é utilizada como representação latente que alimenta os mecanismos de atenção da componente decodificadora, conforme ilustrado na Figura 3.

O modelo baseado na camada codificadora (Figura 1) são comumente empregados em tarefas onde o foco está na compreensão da sequência dos dados de entrada, conforme é observado na arquitetura BERT [Sun et al. 2019, Garrido-Merchan et al. 2023]. Já o modelo construído sobre a camada decodificadora da arquitetura *Transformer* (Figura 2) é empregado em tarefas com características mais distintas, como Reconhecimento de Enlace Textual, Resposta Automática a Perguntas, Similaridade Semântica e, inclusive, Classificação, como observa-se o modelo *GPT* [Radford et al. 2018].

Ainda, é válido ressaltar que os modelos neurais desenvolvidos foram utilizados em dimensões drasticamente reduzidas quando comparados ao artigo original [Vaswani et al. 2017]. O modelo original foi treinado com o auxílio de 8 GPUs do modelo P100, e neste trabalho não foi feito o uso de unidades de processamento gráfico. Para fins comparativos, o modelo original possuía as dimensões de *embedding* de 512 - isto é, cada *token* da entrada é convertido em um vetor de tamanho 512 - e uma rede neural *feedforward* com camada oculta de dimensão 2048, além de contar com oito cabeças codificadoras e decodificadoras. Os modelos deste artigo, por outro lado, foram construídos com dimensões de 64 para *embedding*, 64 para a rede neural *feedforward* e 2 cabeças codificadoras e/ou decodificadoras.

2.4. Treinamento

Cada um dos três modelos propostos foi treinado utilizando 70% da base de dados. Além disso, cerca de 10% dos dados foram reservados para validação durante o treinamento, enquanto os 20% restantes compuseram o conjunto de teste. O processo de treinamento foi realizado em lotes de tamanho 64(*batch size*), por até 10 épocas, com aplicação do critério de *early stopping*, que interrompe o treinamento caso a variação do erro quadrático médio

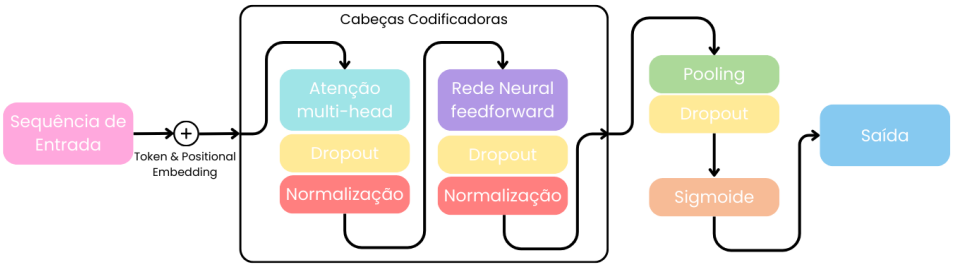

Figura 1. Arquitetura do modelo *encoder-only*. Fonte: Elaboração própria.

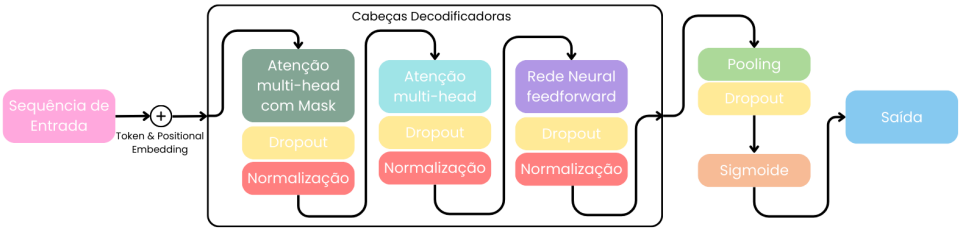

Figura 2. Arquitetura do modelo *decoder-only*. Fonte: Elaboração própria.

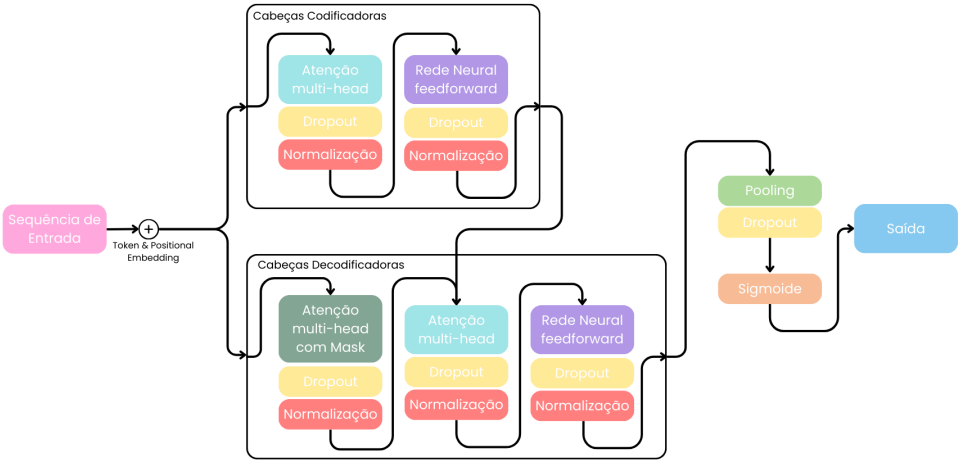

Figura 3. Arquitetura *Transformer* completa. Fonte: Elaboração própria.

entre épocas consecutivas seja inferior a $0,0005$. Também nesta etapa, foi aplicado o *dropout* (conforme ilustram as Figuras 1, 2 e 3), onde adotou-se uma taxa de 50% nas cabeças codificadora e decodificadora e 35% fora delas.

3. Modelos Pré-Treinados

Para avaliar o desempenho dos modelos propostos, realizou-se uma comparação com os seguintes modelos pré-treinados: "*DistilBERT Base*" e "*DistilBERT Portuguese*", dois modelos "*BERT*" [Devlin et al. 2019], além do "*Portuguese T5 Small*". Cada um desses modelos passou por um processo de *fine-tuning*, no qual foi realizado um treinamento reduzido - de até 3 épocas -, utilizando os mesmos dados e *holdout* empregados no treinamento dos modelos propostos, para adequar os pré-treinados à tarefa de classificação de notícias falsas.

O *DistilBERT Base* é um modelo dito *destilado*, isto é, um modelo menor,

mais rápido e leve, treinado para se aproximar do comportamento de um modelo maior (neste caso o *BERT*) [Sanh et al. 2020]. De forma semelhante, o *DistilBERT Portuguese* [Adalberto Ferreira Barbosa Junior 2024] é um modelo destilado a partir de um modelo *BERTimbau* [Souza et al. 2020], que é um modelo *BERT* treinado exclusivamente em português. Já o *Portuguese T5 Small* é um modelo T5 (*Text-to-Text Transfer Transformer*) também treinado na língua Portuguesa [Carmo et al. 2020], que utiliza apenas a camada codificadora, uma vez que o modelo original é direcionado para tarefas de geração de texto (*sequence-to-sequence*).

4. Resultados

Antes de realizar quaisquer comparações acerca dos modelos citados e seus resultados, é importante destacar a diferença entre a quantidade de parâmetros de cada um deles. Como mencionado na Seção 2.3, os modelos desenvolvidos neste artigo foram executados sob dimensões reduzidas, o que não acontece com os modelos pré-treinados que, mesmo optando pelas alternativas destiladas, apresentam uma quantidade de parâmetros significativamente maior.

Os modelos propostos possuem dimensões muito próximas, e cerca de 95% dos seus parâmetros são referentes à camada de *Token And Positional Embedding*. Ao comparar a quantidade de parâmetros do modelo *Transformer* desenvolvido neste artigo com os demais, é possível observar que os modelos *DistilBERT* e *DistilBERT Portuguese* têm 24 vezes mais parâmetros, enquanto o modelo adaptado a partir do codificador do *Portuguese T5* possui cerca de 13 mais parâmetros (conforme mostrado na Tabela 1).

Tabela 1. Tabela comparativa de quantidade de parâmetros e tamanho em disco

	Quantidade de Parâmetros	**Tamanho em Disco**
Encoder-only	2.639.041	10,07 MB
Decoder-only	2,672,385	10,19 MB
Transformer	2,714,177	10,35 MB
DistilBERT	66.362.880	253,15 MB
DistilBERT (PT)	66.395.904	253,28 MB
T5 Encoder (PT)	35.330.816	134,78 MB

Essa diferença tão significativa nas dimensões dos modelos reflete diretamente no tempo total gasto na etapa de treinamento, também denominada *fine-tuning* para os modelos pré-treinados, como pode ser elucidado na Tabela 2. É possível reparar que, mesmo sob uma quantidade superior de épocas, o tempo gasto pelos modelos desenvolvidos foi significativamente inferior, de modo que o modelo que apresentou o menor tempo gasto foi o *Encoder-Only*, mesmo realizando o maior número de épocas (6).

Contudo, esta diferença não foi refletida de maneira diretamente proporcional nas outras métricas observadas: Acurácia, Precisão, *Recall* e *F1-Score*. Como mostra a Tabela 2, o modelo que obteve melhor desempenho foi o *Encoder-Only*, superando todos os outros nas quatro métricas mencionadas. Além disso, os três modelos desenvolvidos e treinados apresentaram resultados superiores aos modelos pré-treinados. Por último, é importante destacar também que os modelos pré-treinados em português mostraram resultados levemente superiores ao modelo pré-treinado em inglês.

Tabela 2. Tabela comparativa de resultados do treinamento dos modelos.

	Tempo	Épocas	Acurácia	Precisão	Recall	F1-Score
Encoder-only	**155.544s**	6	**96,72%**	**96,81%**	**96,72%**	**96,71%**
Decoder-only	245.359s	5	95,05%	95,26%	95,05%	95,05%
Transformer	303.574s	4	95,76%	95,84%	95,76%	95,77%
DistilBERT	3985.493s	3	89,39%	90,13%	89,39%	89,31%
DistilBERT (PT)	4367.343s	3	91,19%	91,30%	91,19%	91,18%
T5 Encoder (PT)	1645.705s	3	91,19%	91,21%	91,19%	91,19%

5. Conclusão

O presente trabalho teve como objetivo implementar e comparar arquiteturas de redes neurais, essencialmente baseadas em redes *Transformers*, na tarefa de classificação de notícias falsas, utilizando duas bases de dados que uniu notícias coletadas de cinco fontes distintas da internet. Foram desenvolvidos e treinados três modelos baseados nos componentes da arquitetura do *Transformer* original: um *Encoder-Only*, um *Decoder-Only* e um *Transformer* (*Encoder-Decoder)*. Além destes, também foram testados os seguintes modelos pré-treinados, submetidos a um *fine-tuning* para adaptá-los à tarefa mencionada: *DistilBERT* (em inglês e em português) e *T5* (do qual utilizou-se apenas a camada codificadora (*Encoder*). A análise comparativa destes envolveu os valores de métricas de avaliação, o tempo total gasto na etapa de treinamento (ou *fine-tuning*) e a dimensão dos modelos em termos de quantidade de parâmetros e tamanho em disco.

Dentre todos os modelos utilizados, os que foram desenvolvidos neste artigo e treinados especificamente para a tarefa de Classificação de Notícias Falsas superaram os modelos pré-treinados com *fine-tuning* em todos os indicadores de desempenho observados, apesar de serem consideravelmente menores e menos complexos. Essa evidência reforça a qualidade do treinamento direcionado para uma tarefa específica, sugerindo que, dependendo das características do problema abordado, arquiteturas mais simples, podem ser mais eficazes do que grandes modelos genéricos pré-treinados.

Já entre os modelos pré-treinados, foi possível observar que os que foram treinados em português (*DistilBERT Portuguese* e *Portuguese T5*) apresentaram desempenho superior ao modelo treinado em inglês (*DistilBERT*), o que reitera a relevância do idioma de pré-treinamento para tarefas situadas no mesmo idioma. Dos dois modelos em português, aquele adaptado a partir da camada codificadora do *T5* foi o que mais se destacou, uma vez que performou extremamente similar ao *DistilBERT Portuguese*, mesmo possuindo menos de metade da quantidade de seus parâmetros e, consequentemente, demandando um tempo significativamente menor na etapa de *fine-tunning*.

Por fim, visando o desenvolvimento de trabalhos futuros, destaca-se a possibilidade de aplicar técnicas de *fine-tuning* mais avançadas, como *Low-Rank Adaptation* (LoRA), ou mesmo estratégias mais específicas, como proposto em [Sun et al. 2019]. Além disso, é importante realizar execuções com conjuntos de dados mais extensos, assim como utilizar arquiteturas de dimensões maiores, uma vez que as limitações de hardware enfrentadas impuseram limitações a este trabalho. Por fim, também é válida a exploração de diferentes arquiteturas e modelos, não apenas restringindo-se aos pré-treinados.

Referências

Adalberto Ferreira Barbosa Junior (2024). distilbert-portuguese-cased (revision df1fa7a).

Cantarino, F. H. S. (2024). Criação de um corpus português para auxiliar a identificação de notícias verdadeiras e falsas. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) – Universidade Federal de Uberlândia.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

G1 - Fato ou Fake (2025). Fato ou fake - o serviço de checagem de fatos do grupo globo. `https://g1.globo.com/fato-ou-fake/`. Acesso em: 7 jun. 2025.

Garrido-Merchan, E. C., Gozalo-Brizuela, R., and Gonzalez-Carvajal, S. (2023). Comparing bert against traditional machine learning models in text classification. *Journal of Computational and Cognitive Engineering*, 2(4):352–356.

IBGE (2023). Pesquisa nacional por amostra de domicílios contínua.

Khyani, D. and B S, S. (2021). An interpretation of lemmatization and stemming in natural language processing. *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology*, 22:350–357.

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*, pages 324–334. Springer International Publishing.

Narde, W., Mendanha, J., Barbosa, H., Coelho, F., Santos, B., and Torres, L. (2024). Classificação de notícias em português utilizando modelos baseados em transferência de aprendizagem e transformers. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 212–216, Porto Alegre, RS, Brasil. SBC.

Pires, V. and e Silva, D. G. (2024). Portuguese fake news classification with bert models. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, pages 834–845, Porto Alegre, RS, Brasil. SBC.

Poynter (2022). A global study on information literacy.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In Sun, M., Huang, X., Ji, H., Liu, Z., and Liu, Y., editors, *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.