Detection of texts generated by LLMs in Portuguese

Guilherme S. M. de C. Paes[1], **Arthur Negrão de F. M. C.**[1], **Guilherme Silva**[1], **Ederson Júnior**[1], **Eduardo Luz**[2], **Pedro Silva**[2]

[1] Postgraduate Program in Computer Science – Federal University of Ouro Preto
35400-000 – Ouro Preto – MG – Brazil

[2]Computing Department – Federal University of Ouro Preto
35400-000 – Ouro Preto – MG – Brazil

{guilherme.paes,arthur.negrao,guilherme.lopes}@aluno.ufop.edu.br
ederson.junior@aluno.ufop.edu.br,{eduluz,silvap}@ufop.edu.br

Abstract. *With the increasing accessibility and use of generative Artificial Intelligence (AI) models, concerns about the misuse of these technologies have intensified. Although originally developed to assist with everyday tasks, their malicious use can contribute to plagiarism and the spread of misinformation. Due to their recent emergence and high capacity, texts generated by Large Language Models (LLMs) still pose significant challenges in terms of detection.*
In this context, this work proposes the construction of a Portuguese-language dataset containing examples of human-authored texts, AI-generated texts, and human texts rewritten by LLMs. Additionally, five classification models were developed based on architectures from the LLaMA and BERT families, along with a recurrent neural network using bidirectional LSTM layers. The proposed classifiers demonstrated strong performance, achieving accuracies of up to 98.18% in binary classification (LLM-authored or not) and 97.7% in the three-class classification task (human, AI-generated, and AI-rewritten), using the defined test set.

Resumo. *Com o crescente acesso e uso de modelos de Inteligência Artificial (IA) generativa, intensificam-se as preocupações quanto ao uso indevido dessas tecnologias. Embora tenham sido desenvolvidos com o objetivo de auxiliar tarefas cotidianas, seu uso mal-intencionado pode favorecer a ocorrência de plágio e a disseminação de desinformação. Devido à sua recente introdução e elevada capacidade, os textos produzidos por Large Language Models (LLMs) ainda apresentam desafios consideráveis em sua detecção.*
Diante desse cenário, este trabalho propõe a construção de um conjunto de dados em língua portuguesa contendo exemplos de textos de autoria humana, textos gerados por IA e textos humanos reescritos por LLMs. Foram também desenvolvidos cinco modelos classificadores baseados em arquiteturas da família LLaMA e BERT, além de uma rede neural recorrente com camadas LSTM bidirecionais. Os classificadores propostos apresentaram desempenho expressivo, alcançando acurácias de até 98,18% na classificação binária (texto escrito ou não por LLM) e 97,7% na classificação de três classes (humano, gerado por IA e reescrito por IA), utilizando o conjunto de teste definido.

1. Introdução

Nos últimos anos, observou-se um avanço significativo no desenvolvimento de modelos de Inteligência Artificial (IA) generativa, como o ChatGPT, DeepSeek e Gemini. Os modelos atuais, denominados *Large Language Models* (LLMs), distinguem-se das abordagens tradicionais, como as *Recurrent Neural Networks* (RNNs), tanto pela expressiva quantidade de parâmetros quanto pela capacidade de compreensão e geração em uma ampla gama de tarefas. Entre essas tarefas, destacam-se a resposta a perguntas, a geração de textos, o reconhecimento e a produção de mídias diversas (imagens, vídeos e áudios), bem como a geração de código em múltiplas linguagens [Gemini Team Group 2024].

Além da melhoria no desempenho desses modelos, destaca-se também, nos últimos anos, a disponibilização pública de ferramentas baseadas em LLMs. Essas ferramentas oferecem à população recursos acessíveis para automatizar e aprimorar atividades cotidianas, como a redação de textos extensos [Rossoni and Chat 2022] e o apoio a processos educacionais [Sant et al. 2023], por meio da geração, revisão e interpretação de conteúdos textuais e visuais.

Embora essa mudança de paradigma seja promissora e traga inovações relevantes, seu uso inadequado pode acarretar problemas significativos, como a ocorrência de plágio não identificado e, o mais preocupante, a disseminação de informações falsas por meio de notícias geradas parcial ou integralmente por IA [Pires et al. 2024, Gôlo et al. 2024]. O risco central reside na possibilidade de que informações subsequentes se fundamentem nesses textos artificiais, perpetuando informações incorretas [Else 2023], o que pode gerar impactos negativos especialmente em áreas sensíveis como a medicina [DONATO et al. 2023], nas quais a veracidade dos dados é crucial para a saúde e segurança das pessoas. Torna-se, portanto, evidente a necessidade de avanços científicos no desenvolvimento de métodos eficazes para detectar o uso indevido e mal-intencionado dessas ferramentas.

Um fator adicional que agrava essa problemática é a escassez de estudos voltados à detecção de textos artificiais em língua portuguesa, diferente do que ocorre para a língua inglesa [Wu et al. 2023]. Questões socioeconômicas, aliadas à complexidade gramatical do idioma, estão entre os principais obstáculos para o avanço dessa linha de pesquisa [Caseli and Nunes 2023]. Considerando que o português é a quarta língua mais falada no mundo [instituto camoes 2022], é essencial fomentar o desenvolvimento de conjuntos de dados e classificadores especificamente voltados para essa língua.

Dada a recente e rápida evolução dos LLMs, a tarefa de classificar textos por eles gerados apresenta desafios consideráveis, como superar a elevada performance desses modelos em tarefas linguísticas e contornar a escassez de dados que capturem adequadamente suas características textuais. No entanto, considerando a alta capacidade dos próprios LLMs em diversas tarefas de Processamento de Linguagem Natural (PLN), supõe-se que esses modelos possam ser eficazes também na detecção de textos produzidos por outros LLMs.

As principais contribuições deste trabalho são: (1) A introdução do PT-Detect, um novo conjunto de dados para detecção de texto em português gerados por IA; (2) Uma análise comparativa de seis modelos com arquiteturas variadas, desde RNNs até, modelos mais complexos como BERT e LLMs; (3) A demonstração de que traços linguísticos per-

manecem detectáveis mesmo em textos parafraseados por LLMs. Os modelos alcançaram acurácias de 98,18% e 97,7% nas tarefas de classificação binária (texto escrito ou não por LLM) e ternária (texto escrito, não escrito e reescrito por LLM), respectivamente.

O presente trabalho é organizado de acordo com a seguinte estrutura: a Seção 2 aborda uma revisão da literatura. Já a Seção 3 define a metodologia que será aplicada para o desenvolvimento do trabalho. Enquanto a Seção 4 traz os resultados obtidos após o desenvolvimento do trabalho e uma discussão sobre sua relevância. Por fim, a Seção 5 conclui o trabalho apresentado.

2. Trabalhos Relacionados

Embora o crescente avanço das técnicas de PLN, observa-se uma lacuna significativa no desenvolvimento de estudos voltados à língua portuguesa. Com cerca de 260 milhões de falantes em 2022, sendo o quarto idioma mais falado do mundo [instituto camoes 2022], o português ainda é pouco contemplado em estratégias que considerem suas especificidades linguísticas.

Historicamente, fatores socioeconômicos direcionaram a maior parte das pesquisas em PLN para a língua inglesa [Caseli and Nunes 2023]. No entanto, o português impõe desafios adicionais, como alta complexidade gramatical e grande variação dialetal [Pinto 2012, Castro 2006], o que dificulta a criação de modelos generalistas eficazes.

Ainda assim, há avanços relevantes, como os modelos BERTimbau [Souza et al. 2020] e Sabiá [Almeida et al. 2024], treinados especificamente para o português. Esses modelos superam abordagens generalistas em tarefas como a detecção de discurso de ódio [da Silva Oliveira et al. 2024, Leite et al. 2020], evidenciando os benefícios de adaptações linguísticas específicas.

Modelos de linguagem de larga escala (LLMs) destacam-se por seu desempenho em tarefas como geração de texto, código e classificação [Chakraborty et al. 2023, Liu et al. 2024, Zhang et al. 2024], mas ainda apresentam limitações ao mimetizar completamente a escrita humana. Textos gerados por LLMs tendem a ser mais longos, menos variados lexicalmente e semanticamente mais neutros, reflexo de seus processos de treinamento e mitigação de vieses [Soni and Wade 2023, Wu et al. 2023].

Diante disso, cresce o interesse por métodos de detecção de textos artificiais, utilizando técnicas como watermarking, análise estatística e os próprios LLMs como classificadores [Wu et al. 2023]. No entanto, a maioria dos estudos concentra-se em inglês e chinês, resultando em escassez de soluções voltadas ao português.

O conjunto de dados HC3 [Guo et al. 2023] é focado em dois idiomas: inglês e chinês. Ele revela padrões comuns em textos humanos e artificiais desses dois idiomas, como maior concisão e diversidade lexical nos primeiros. Tais padrões podem ser indicativos úteis para a detecção automática de textos gerados por IA em diversas línguas. Diferente do HC3, este trabalho propõe a construção de uma base de dados específica do Português, gerada com um modelo especializado também no Português.

Assim, apesar de serem capazes de produzir textos de alta complexidade e difícil diferenciação, os LLM ainda são marcados por características específicas e compartilhadas entre diferentes línguas que podem ser exploradas para uma classificação de textos gerados por esses modelos.

3. Metodologia Proposta

Esta seção apresenta a metodologia seguida para o desenvolvimento do trabalho, desde a construção do conjunto de dados proposto, a partir do uso de um conjunto externo de textos jornalísticos escritos por humanos, além da descrição dos métodos para construção, treinamento e avaliação dos modelos classificadores.

3.1. Proposição do Conjunto de Dados

Para a construção do conjunto de dados proposto, utilizou-se como base o *News of the Brazilian Newspaper* [Marlesson 2024], que contém aproximadamente 160 mil notícias extraídas do portal Folha de São Paulo, publicadas entre janeiro de 2015 e setembro de 2017. A escolha desse conjunto se deve à sua diversidade temática e ao fato de representar um período anterior à popularização dos LLMs, assegurando a autenticidade dos textos humanos.

O novo conjunto inclui três categorias de texto: (i) notícias originais escritas por humanos; (ii) versões inteiramente geradas por modelos de linguagem; e (iii) versões reescritas por LLMs a partir das notícias originais. Os textos artificiais foram produzidos com base em *prompts* cuidadosamente elaborados: para a geração, solicitou-se ao modelo criar um texto com base no título original e com tamanho semelhante ao da notícia humana; para a reescrita, o *prompt* foi formulado de modo a preservar o conteúdo, mas utilizando vocabulário distinto. A Tabela 1 apresenta os prompts utilizados.

Tabela 1. *Prompts* usados para geração e reescrita das notícias.

Tarefa	***Prompt* usado**
Geração de notícia	Crie uma notícia jornalística com exatamente *length*(*notícia*) caracteres, incluindo espaços, com o título: *título*. Assegure que o conteúdo seja informativo e esteja dentro do limite de caracteres especificado.
Reescrita de notícia	Reescreva a notícia jornalística abaixo, mantendo o mesmo contexto e informações, mas utilizando suas próprias palavras: *notícia*.

A geração dos textos foi realizada por meio da API da OpenAI, permitindo o ajuste de parâmetros como temperatura (definida em 0,7 para balancear diversidade e coerência) e limite de *tokens* (10.000, para evitar cortes e enviesamento por tamanho). Além dos modelos da OpenAI, foram utilizados modelos da família Sabiá (especificamente o sabia-3), desenvolvidos pela empresa Maritalk, por se tratarem dos LLMs mais avançados treinados exclusivamente em português.

A partir do corpus original, foram selecionados aleatoriamente 1.008 pares de título e notícia para gerar os trios de texto. Esse procedimento assegura representatividade estatística e diversidade temática. Após a geração, todos os textos foram padronizados, com a remoção de caracteres especiais e formatações como a repetição do título no corpo do texto, comum nas respostas do modelo Sabiá.

O conjunto de dados final (disponível por meio da plataforma *Dropbox*[1]) contém 3.024 exemplos, distribuídos igualmente entre as três classes. Os dados foram embaralhados e divididos aleatoriamente em subconjuntos de treino (80%) e teste (20%). Adicionalmente, 10% do conjunto de treino foi reservado para validação durante o treinamento dos modelos. A Tabela 2 apresenta a distribuição final dos exemplos.

[1] `https://www.dropbox.com/scl/fo/rzrcithhjaim3pcck304f/ALopAH9I_ULdfZRUN9kow_s?rlkey=n9cntce46mlfuk1h19tf8yhs5&st=bnsml9ti&dl=0`

Tabela 2. Distribuição do exemplos humanos e artificiais entre os conjuntos de treino e teste.

Classe dos exemplos	Treino	Validação	Teste	Total
Número de exemplos humanos	727	79	202	1008
Número de exemplos artificiais	725	79	204	1008
Número de exemplos reescritos	730	79	199	1008

3.2. Construção e Avaliação dos Classificadores

Como abordagem *baseline*, propôs-se um modelo baseado em RNN, utilizando duas camadas LSTM bidirecionais com 256 e 128 unidades, respectivamente, precedidas por uma camada de *embedding*. O uso de LSTM bidirecional visa capturar o contexto sequencial completo. Aplicou-se *dropout* de 20% após ambas as camadas de LSTM bidirecional para reduzir o risco de *overfitting*. A rede conta ainda com uma camada densa intermediária (32 neurônios e ativação ReLU) e uma camada final com ativação sigmoide (para classificação binária) ou softmax (para classificação com 3 classes).

Para o modelo *baseline*, os dados textuais foram *tokenizados* e padronizados via *padding*. Os *embeddings* foram configurados dinamicamente conforme o vocabulário e o comprimento máximo das sequências. Utilizou-se o *checkpoint* com melhor desempenho em validação, evitando sobreajuste.

Tendo o modelo *baseline*, foram avaliados quatro outros modelos: BERTimbau [Souza et al. 2020], BERTuguês [Zago and Pedotti 2024], multi-BERT [Devlin et al. 2018] (estado da arte para diversas tarefas de PLN em português) e modelos da família LLaMA [Llama Team Group 2024]. A escolha contempla tanto modelos de linguagem treinados especificamente para o português (BERTimbau e BERTuguês) um modelo multilíngue (multi-BERT) e modelos de LLM (LLaMA), permitindo comparações em termos de desempenho e complexidade. A Tabela 3 apresenta o número de parâmetros e o tempo de inferência desses modelos.

Tabela 3. Comparação entre os modelos usados para classificação.

Modelo	Número de parâmetros	Tempo para inferência em GPU (s)
neuralmind/bert-large-portuguese-cased	335M	13,16
ricardoz/BERTugues-base-portuguese-cased	110M	4,45
google-bert/bert-base-multilingual-cased	179M	4,47
meta-llama/Llama-3.2-3B-Instruct	3,21B	229,21
meta-llama/Llama-3.1-8B-Instruct	8,03B	254,73
LSTM proposto	17,8M	2,6

Todos os modelos foram avaliados em duas tarefas: classificação binária (textos humanos vs. textos com interferência artificial) e classificação com três classes (humano, gerado, reescrito). Para os modelos baseados em BERT, foi realizado ajuste fino supervisionado com *labels* correspondentes às classes. Utilizou-se o *tokenizador* nativo de cada modelo.

Os modelos LLaMA avaliados foram o LLaMA 3.1-8B-Instruct e o LLaMA 3.2-3B-Instruct. A comparação entre eles visa analisar o impacto do número de parâmetros frente à atualização do modelo. Devido ao custo computacional elevado, utilizou-se o método LoRA para ajuste fino eficiente. O treinamento foi supervisionado, com passagem

única sobre os dados para evitar sobreajuste, e treinado com LoRA, especificamente com *Rank* e *Alpha* iguais a 64. As instruções foram fornecidas por meio de *prompts* em estilo de pergunta-resposta, conforme segue:

***Prompt 1*:** *"Me diga, com sim ou não, se você considera o seguinte texto como sendo de autoria humana:* `[TEXTO]`*."*

***Prompt 2*:** *"Me diga, com 0, 1 ou 2, se você considera o seguinte texto como sendo de autoria humana, artificial ou reescrito artificialmente:* `[TEXTO]`*."*

Os modelos da família *LLaMA* foram avaliados sob duas abordagens distintas: (i) em regime de *zero-shot* e (ii) com ajuste fino supervisionado (*fine-tuning*). Na abordagem *zero-shot*, os modelos foram utilizados em sua forma original, sem qualquer processo de especialização ou treinamento adicional, realizando a tarefa de classificação apenas com base nos *prompts* previamente definidos.

Na segunda abordagem, os modelos passaram por um processo de ajuste fino com os dados de treinamento do conjunto proposto. Para essa etapa, foi utilizado o método *QLoRA*, que permite a adaptação de modelos de grande porte de forma eficiente em termos computacionais. Após o treinamento, os modelos ajustados foram empregados na tarefa de classificação, possibilitando a comparação de desempenho entre as duas estratégias.

3.3. Métricas de Avaliação

Para avaliar o desempenho dos modelos propostos, foi adotada a métrica de acurácia sobre o conjunto de teste. Essa métrica foi escolhida por ser amplamente utilizada em tarefas de classificação e por apresentar resultados confiáveis em cenários com dados balanceados, como é o caso deste trabalho. A acurácia é definida conforme a Equação (1), considerando os seguintes elementos: (i) ***True Positive (TP)*:** textos corretamente classificados como de autoria humana; (ii) ***True Negative (TN)*:** textos corretamente classificados como de autoria artificial; (iii) ***False Positive (FP)*:** textos artificiais classificados incorretamente como humanos; e (iv) ***False Negative (FN)*:** textos humanos classificados incorretamente como artificiais.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

Além da acurácia, foram utilizadas as métricas de precisão, revocação e F1-score, com o objetivo de fornecer uma avaliação mais abrangente do comportamento dos modelos, especialmente em relação ao equilíbrio entre as classes. Essas métricas são definidas a seguir:

$$\text{Revocação} = \frac{TP}{TP + FN}, \tag{2}$$

$$\text{Precisão} = \frac{TP}{TP + FP}, \tag{3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}. \tag{4}$$

Essas métricas complementares possibilitam a análise da performance dos modelos sob diferentes perspectivas, permitindo a identificação de possíveis vieses ou desequilíbrios nas classificações.

4. Experimentos e Resultados

Nesta seção, são apresentados e analisados os resultados do trabalho de acordo com o que se propõe.

4.1. *Setup* Experimental

A implementação deste trabalho foi realizada utilizando a linguagem de programação *Python*. A construção do conjunto de dados proposto foi feita por meio da API da empresa *OpenAI*, utilizada para a geração e reescrita automática de textos.

A implementação dos classificadores baseados em modelos de linguagem, com ajuste fino supervisionado, foi conduzida com o uso do *framework PyTorch*. Para os modelos LLM da família LLaMA, empregou-se a biblioteca *LlamaFactory*, enquanto os modelos baseados em BERT foram implementados com a biblioteca *Transformers*, da Hugging Face. A avaliação de desempenho dos modelos de linguagem foi realizada com o apoio da biblioteca *scikit-learn*. Já a implementação e avaliação do modelo *baseline* foram conduzidas integralmente com o *framework TensorFlow*. Todas as execuções, tanto para os modelos de linguagem quanto para a RNN, foram realizadas em *notebooks* Jupyter hospedados na plataforma *Google Colab*, utilizando uma GPU A100. O código empregado nesta pesquisa pode ser encontrado publicamente através de seu repositório *Github*. [2]

O treinamento da RNN, considerada como modelo *baseline*, foi feito com uma taxa de aprendizado (*learning rate*) de 10^{-3}, utilizando o otimizador *Adam*. A função de perda adotada foi a *binary crossentropy* para os experimentos de classificação binária e a *categorical crossentropy* para a tarefa com três classes. Os modelos foram treinados por 20 épocas, com salvamento de *checkpoints* a cada cinco épocas. A avaliação final no conjunto de teste utilizou o *checkpoint* que apresentou a menor perda no conjunto de validação, de modo a evitar sobreajuste. Para os modelos de linguagem, o treinamento foi realizado com uso de *data collator* para preparação dos lotes, ao longo de cinco épocas. A taxa de aprendizado adotada foi de 2×10^{-5}, com decaimento de peso (*weight decay*) de 0,01.

4.2. Análise dos Resultados Obtidos

A seguir serão apresentadas as análises e interpretações dos resultados obtidos tanto para o problema binário quanto o de três classes.

4.2.1. Análise dos Dados

O conjunto de dados resultante foi analisado, principalmente, sob a perspectiva da disparidade nos tamanhos dos textos de autoria humana e daqueles gerados por inteligência artificial. Embora o *prompt* utilizado para a geração de textos tenha sido projetado para manter a similaridade de comprimento em relação ao texto original, observou-se uma variação significativa. Ao analisar pares de textos com o mesmo título, verificou-se que em 145 casos os textos humanos apresentaram maior número de caracteres, enquanto em 55 pares os textos gerados pelo modelo Sabiá-3 foram mais extensos.

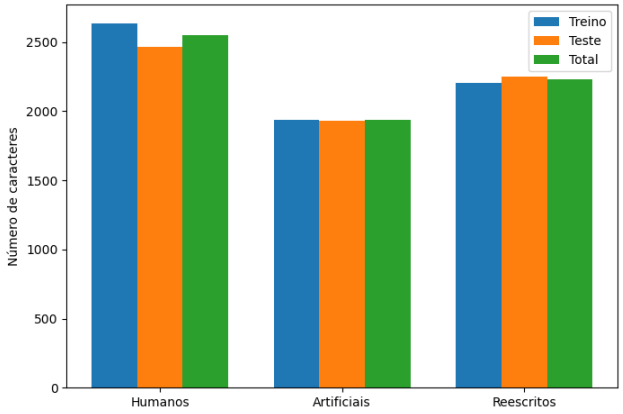
Além da contagem de casos com textos maiores, foi realizada uma análise quantitativa e descritiva dos tamanhos dos textos em cada um dos *subsets* (treino, teste e validação),

[2]Disponível em `https://github.com/GSalimp/PTDetect`.

bem como no conjunto total. A Tabela 2 apresenta a distribuição de amostras por classe em cada *subset*.

A Figura 1 mostra os valores médios de comprimento, em caracteres, para cada classe, e, conforme evidenciado na mesma, os textos humanos apresentaram, em média, comprimento 30% superior aos textos gerados por IA e 15% maior que os textos reescritos. Essa diferença pode introduzir um viés nos modelos classificadores, favorecendo a classificação de textos mais longos como humanos e, inversamente, textos mais curtos como artificiais. No entanto, como os textos reescritos apresentaram comprimento médio intermediário entre os textos humanos e artificiais, não se observa um padrão de viés claro nesta classe.

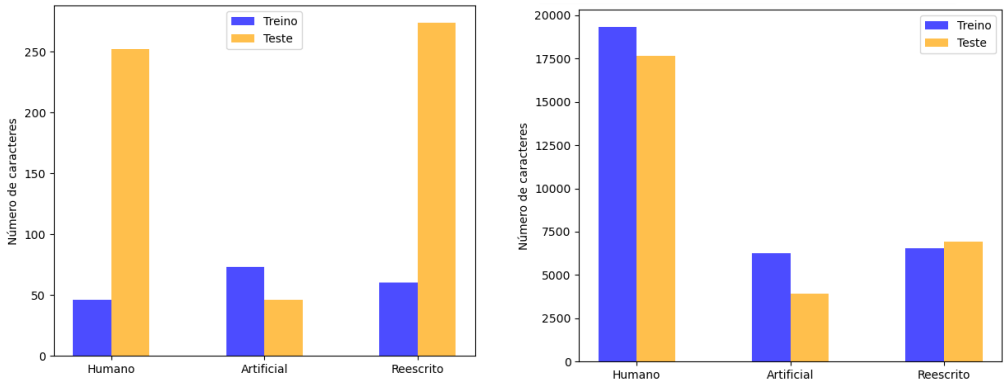
Figura 1. Médias dos tamanhos de texto de cada classe e do conjunto total.



A Figura 2 ilustra os menores e maiores valores observados em cada divisão. Os textos do conjunto de validação foram analisados em conjunto com o de treino, uma vez que a divisão entre ambos foi realizada de forma aleatória durante o treinamento da RNN, impedindo uma análise separada por classe. É possível perceber que os menores textos de cada classe possuem tamanhos similares, sendo que o menor texto humano no conjunto de treino é, inclusive, mais curto que os exemplos mínimos das outras duas classes. Essa distribuição mais ampla permite avaliar, nos experimentos, se o tamanho do texto é, de fato, um fator determinante para a classificação, ou se os modelos são capazes de identificar padrões mais complexos entre as classes. Apesar da predominância de textos humanos mais longos, todas as classes apresentam exemplos distribuídos ao longo de uma faixa ampla de tamanhos, o que contribui para a robustez da análise.

Além da análise dos comprimentos, a Tabela 2 também permite observar que, mesmo com a seleção aleatória dos exemplos humanos e a divisão randômica dos dados entre treino e teste, as quantidades de textos artificiais e reescritos foram praticamente equivalentes nos dois *subsets*. Verifica-se, ainda, uma alta consistência entre os valores estatísticos dos subconjuntos quando comparados ao conjunto total. Os valores médios, em particular, mantêm uma proporção similar entre os conjuntos de treino, teste e geral. Assim, é possível afirmar que o conjunto de dados encontra-se balanceado, tanto em termos de distribuição entre classes quanto de comprimento textual.

Figura 2. Análise sobre o conjunto de dados gerados. (a) apresenta os menores valores de cada classe em cada conjunto, enquanto (b), os maiores.



(a) (b)

4.2.2. Resultados dos Classificadores

A Tabela 4 apresenta os resultados obtidos para as tarefas de classificação binária e ternária (três classes), tanto no cenário *zero-shot* quanto após o ajuste fino dos modelos.

Tabela 4. Resultados da classificação em duas e três classes com (✓) e sem *0-Shot* (×).

Métodos	***Zero-Shot***	**Acurácia (%)**	**Precisão (%)**	**Revocação (%)**	**F1-score**
Classificação Binária (texto original e gerado por IA)					
LLM Multi-BERT	✓	66,4	44,3	66,4	0,532
LLM BERTugues	✓	61,0	61,6	61,0	0,613
LLM BERTimbau-Large	✓	60,5	57,8	60,5	0,591
RNN LSTM Bidirecional	✓	57,0	48,0	48,0	0,480
LLM Llama 3.1-8B	✓	49,6	46,6	49,6	0,480
LLM Llama 3.2-3B	✓	44,4	49,5	49,5	0,495
LLM Llama 3.1-8B	×	**98,2**	97,5	**98,5**	**0,980**
LLM BERTimbau-Large	×	97,7	**97,7**	97,7	0,977
LLM Llama 3.2-3B	×	97,7	97,4	97,4	0,974
LLM BERTugues	×	94,7	94,9	94,7	0,948
LLM Multi-BERT	×	85,5	86,9	85,5	0,862
RNN LSTM Bidirecional	×	73,0	72,0	63,0	0,672
Classificação das três classes (texto original, reescrito e gerado por IA)					
LLM Llama 3.1-8B	✓	36,0	41,0	36,0	0,383
RNN LSTM Bidirecional	✓	35,0	28,0	35,0	0,311
LLM BERTimbau-Large	✓	34,2	25,5	34,2	0,292
LLM Multi-BERT	✓	32,7	20,4	32,7	0,211
LLM Llama 3.2-3B	✓	31,9	32,2	31,7	0.319
LLM BERTugues	✓	30,6	16,6	30,6	0,215
LLM Llama 3.1-8B	×	**97,7**	**97,7**	**97,7**	**0,977**
LLM BERTimbau-Large	×	96,5	96,5	96,5	0,965
LLM Llama 3.2-3B	×	93,0	93,0	93,0	0,930
LLM BERTugues	×	91,9	92,1	91,9	0,919
LLM Multi-BERT	×	82,0	85,8	82,0	0,839
RNN LSTM Bidirecional	×	56,0	58,0	56,0	0,560

A análise dos resultados obtidos nas classificações em regime *zero-shot* oferece insights relevantes sobre as características do conjunto de dados e a complexidade da tarefa. O conjunto de teste encontra-se razoavelmente balanceado para a tarefa de três classes e apresenta uma proporção de aproximadamente 2:1 entre exemplos artificiais e

humanos na tarefa binária, uma vez que as classes “gerado” e “reescrito” são agrupadas nessa configuração.

Nessas condições, as médias de acurácia obtidas pelos modelos em *zero-shot* variaram entre 30% e 36% na tarefa de três classes e entre 44% e 65% na tarefa binária (Tabela 4). Esses resultados indicam que os modelos, sem especialização prévia, apresentam desempenho próximo ao esperado por uma classificação aleatória, sugerindo que padrões discriminativos não são facilmente extraídos sem treinamento. Assim, os ganhos observados com o ajuste fino podem ser atribuídos à capacidade de aprendizado dos modelos, e não a um viés intrínseco nos dados.

Entre os modelos ajustados, o LLaMA 3.1 8B apresentou o melhor desempenho, alcançando acurácias de 98,18% na tarefa binária e 97,7% na tarefa de três classes. Os modelos BERTimbau, LLaMA 3.2 3B e BERTuguês também obtiveram resultados expressivos, com acurácias superiores a 90% em ambas as configurações. Em seguida, o modelo Multi-BERT apresentou desempenho inferior, porém ainda aceitável. Por outro lado, o modelo baseado em LSTM apresentou acurácia consideravelmente mais baixa, evidenciando as limitações de arquiteturas recorrentes em tarefas que requerem alto nível de abstração textual, como as aqui propostas, nas quais modelos baseados em *transformers* demonstram maior eficácia.

Observa-se, ainda, que todos os modelos obtiveram melhores desempenhos na classificação binária em comparação à classificação ternária. Esse resultado corrobora a hipótese de que a introdução de uma terceira classe , textos reescritos por IA , aumenta a complexidade da tarefa. No entanto, os bons resultados obtidos indicam que essa classe intermediária também apresenta padrões linguísticos identificáveis, que podem ser utilizados pelos modelos para fins de classificação.

Embora o modelo LLaMA 3.1 8B tenha se destacado em termos absolutos, os modelos BERTimbau e BERTuguês, ambos treinados especificamente em português, apresentaram desempenho competitivo, superando inclusive o modelo LLaMA 3.2 3B, que possui aproximadamente 2,9 bilhões de parâmetros a mais (conforme Tabela 3). Quando comparados ao modelo Multi-BERT, os modelos especializados para a língua portuguesa também demonstraram desempenho significativamente superior. Essa diferença reforça a hipótese de que modelos treinados com foco em uma tarefa ou idioma específico tendem a superar modelos generalistas, especialmente quando estes últimos demandam maior capacidade computacional para atingir desempenhos semelhantes.

Diante dos resultados apresentados, é possível concluir que a tarefa de classificação de textos gerados por LLMs, incluindo textos parafraseados por IA, é viável e apresenta desempenho satisfatório. Além disso, os textos reescritos por modelos de linguagem mantêm traços estruturais detectáveis, reforçando a hipótese central deste trabalho.

5. Considerações Finais

Este trabalho propôs um conjunto de dados em português para a detecção de textos gerados e reescritos por LLMs, além do desenvolvimento e avaliação de diferentes modelos classificadores. Os resultados obtidos (com acurácias superiores a 97%) demonstram que há padrões linguísticos distinguíveis entre textos humanos e artificiais, inclusive na categoria de textos reescritos, que configuraram-se como uma tarefa mais complexa. Modelos treinados especificamente para o português apresentaram desempenho competitivo em relação

a modelos multilíngues mais robustos, evidenciando a eficácia de soluções linguísticas especializadas.

Considerando os problemas encontrados nos dados propostos, fica como trabalho futuro a criação do conjunto de dados proposto com todos os exemplos do conjunto *News of the Brazilian Newspaper*, o qual ainda não foi feito por questões de limitações computacionais. Para avaliar ainda melhor a nova metodologia, trabalhos futuros devem também se concentrar em aplicar diferentes perspectivas de testes sobre os algoritmos propostos, como aumentar a base de dados e testar modelos de linguagem mais complexos.

Agradecimentos

Os autores gostariam de agradecer à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, subsídio APQ-01768-24), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Universidade Federal de Ouro Preto (UFOP/PROPPI) por apoiar o desenvolvimento do presente estudo.

Referências

Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models.

Caseli, H. d. M. and Nunes, M. d. G. V. (2023). Processamento de linguagem natural: conceitos, técnicas e aplicações em português.

Castro, I. (2006). Introdução à história do português.

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., and Huang, F. (2023). On the possibilities of ai-generated text detection.

da Silva Oliveira, A., de Carvalho Cecote, T., Alvarenga, J. P. R., da Silva Luz, E. J., et al. (2024). Toxic speech detection in portuguese: A comparative study of large language models. In Proceedings of the 16th International Conference on Computational Processing of Portuguese, pages 108–116.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

DONATO, H., ESCADA, P., and VILLANUEVA, T. (2023). A transparência da ciência com o chatgpt e as ferramentas emergentes de inteligência artificial: como se devem posicionar as revistas científicas médicas. The Transparency of Science with ChatGpt and the Emerging Artificial Intelligence Language Models: Where Should Medical Journals Stand.

Else, H. (2023). Abstracts written by chatgpt fool scientists. Nature, 613(7944):423–423.

Gemini Team Group (2024). Gemini: A family of highly capable multimodal models.

Gôlo, M. P. S., Mori, A. L. V., Oliveira, W. G., Barbosa, J. R., Graciano Neto, V. V., Lima, E. A. d., and Marcacini, R. M. (2024). On the use of large language models to detect brazilian politics fake news. In Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). SBC.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., and Wu, Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

instituto camoes (2022). Dia mundial da lingua portuguesa: 5 de maio de 2022. Acessado em 05 de junho de 2024.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Wong, K.-F., Knight, K., and Wu, H., editors, Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Liu, F., Liu, Y., Shi, L., Huang, H., Wang, R., Yang, Z., Zhang, L., Li, Z., and Ma, Y. (2024). Exploring and evaluating hallucinations in llm-powered code generation.

Llama Team Group (2024). The llama 3 herd of models.

Marlesson (2024). News of the brazilian newspaper. Kaggle. Accessado em 31-08-2024.

Pinto, J. (2012). A aquisição de português le por alunos marroquinos: dificuldades interlinguísticas. In Actas del II Congreso Internacional de la Sociedad Extremeña de Estudios Portugueses y la Lusofonía (SEEPLU), pages 217–239. SEEPLU-CILEM-LEPOLL.

Pires, V. B., Guerreiro, D., et al. (2024). Portuguese fake news classification with bert models. In Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), pages 834–845. SBC.

Rossoni, L. and Chat, G. (2022). A inteligência artificial e eu: escrevendo o editorial juntamente com o chatgpt. Revista Eletrônica de Ciência Administrativa, 21(3):399–405.

Sant, F. P., Sant, I. P., de Camargo Sant, C., et al. (2023). Uma utilização do chat gpt no ensino. Com a Palavra, o Professor, 8(20):74–86.

Soni, M. and Wade, V. (2023). Comparing abstractive summaries generated by chatgpt to real summaries through blinded reviewers and text classification algorithms. arXiv preprint arXiv:2303.17650.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9, pages 403–417. Springer.

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., and Chao, L. S. (2023). A survey on llm-gernerated text detection: Necessity, methods, and future directions. arXiv preprint arXiv:2310.14724.

Zago, R. and Pedotti, L. d. S. (2024). Bertugues: A novel bert transformer model pre-trained for brazilian portuguese.

Zhang, Y., Wang, M., Ren, C., Li, Q., Tiwari, P., Wang, B., and Qin, J. (2024). Pushing the limit of llm capacity for text classification.