Uma Avaliação Comparativa entre o DepreBERTBR e Modelos de Linguagem para Classificação de Textos Depressivos

Ayrton Douglas Rodrigues Herculano[1], Laerty Santos da Silva[1], Damires Yluska de Souza Fernandes[1], Alex Sandro da Cunha Rego[1]

[1]Instituto Federal da Paraíba (IFPB)
Av. 1º de Maio, 720 – Jaguaribe – João Pessoa – PB – Brasil

{ayrton.herculano,laerty.santos}@academico.ifpb.edu.br, {damires,alex}@ifpb.edu.br

Abstract. *Depression has incapacitated people around the world. Identifying signs of this disorder in textual reports, in environments where an individual feels more comfortable, may enable early medical interventions. The DepreBERTBR is a pre-trained language model with specialized content in the domain of depression, likely to be explored. This work presents an experimental comparative evaluation between DepreBERTBR and general-domain monolingual or multilingual language models, considering Brazilian Portuguese and the task of classifying depression into four classes (absent, mild, moderate, or severe). The results show that DepreBERTBR is competitive, matching other models in terms of F1-score.*

Resumo. *A depressão tem incapacitado pessoas mundialmente. Identificar sinais desse transtorno em relatos textuais, em ambientes onde um indivíduo se sente mais à vontade, pode viabilizar intervenções médicas. O DepreBERTBR é um modelo de linguagem pré-treinado com conteúdo especializado no domínio da depressão, com potencial para ser explorado. Este trabalho apresenta uma avaliação experimental comparativa entre o DepreBERTBR e modelos de linguagem de domínio geral monolíngue e multilíngue, considerando o português brasileiro e a tarefa de classificação da depressão em quatro classes (ausente, leve, moderada ou grave). Os resultados mostram que o DepreBERTBR é competitivo, equiparando-se em termos de F1-score aos outros modelos.*

1. Introdução

A depressão é um transtorno mental que tem acometido milhões de pessoas em todo o mundo. No Brasil, sua incidência vem aumentando e desestabilizando a rotina social e profissional de pessoas acometidas pela doença [OMS 2023]. Sintomas tais como baixa autoestima, sensação de culpa, desânimo e solidão são característicos na depressão e podem evoluir, em casos extremos, para o cometimento de suicídio [American Psychiatric Association 2023]. Neste cenário sensível e grave, iniciativas que auxiliem na identificação precoce de sinais da depressão têm sido cada vez mais buscadas. Ações nesse sentido podem ser cruciais para viabilizar intervenções médicas e/ou psicológicas oportunas, prevenindo o agravamento de sintomas da doença. Uma abordagem promissora para identificar indícios de depressão consiste na análise de relatos textuais espontâneos compartilhados pelos próprios indivíduos em redes sociais ou em

aplicações voltadas para o monitoramento do humor, a exemplo do Woebot e do Sanvello [Pavlopoulos et al. 2024].

Como ilustração, a rede social Reddit[1] tem se consolidado como um fórum online onde usuários compartilham relatos e opiniões por meio de interações em comunidades temáticas denominadas *subreddits*. Nessas comunidades, os participantes normalmente expressam seus pensamentos, sentimentos, medos e experiências particulares de forma aberta. A manifestação de tais emoções e vulnerabilidades demonstra que muitos usuários sentem-se mais à vontade em externalizar suas preocupações nesses ambientes do que em contextos clínicos tradicionais, como os consultórios médicos [Uban et al. 2021]. Dessa forma, pesquisas têm sido realizadas com o objetivo de compreender, por exemplo, a relação entre relatos postados na Reddit e a suposição de um nível de severidade de depressão [Naseem et al. 2022, Poświata and Perełkiewicz 2022]. Estudos dessa natureza integram conhecimentos oriundos das áreas de Psiquiatria, Psicologia, Sociolinguística e Neurociência [Ji et al. 2022, Costa et al. 2023], aliados ao emprego de técnicas computacionais como Análise de Sentimentos, Processamento de Linguagem Natural (PLN) e Modelos de Linguagem de Grande Escala (*Large Language Model* - LLM).

Particularmente, os LLMs (modelos de linguagem neural, caracterizados como método de IA generativa e treinados a partir de grandes volumes de dados textuais), tornaram-se ferramentas cada vez mais relevantes para a automatização de tarefas tais como sumarização, tradução, classificação de texto, entre outros [Caseli and Nunes 2024]. No entanto, a eficácia desses modelos está diretamente relacionada à sua capacidade de compreender e interpretar com precisão as complexidades linguísticas, dentro do contexto de um idioma, o que inclui suas nuances semânticas, vocabulário e terminologias específicas dependentes do domínio em que são aplicados [Silva et al. 2024]. Como ilustração, LLMs podem ser usados, na área da Saúde, para apoiar tarefas de sumarização de evoluções clínicas [Caseli and Nunes 2024].

Em se tratando de tarefas como a predição na área da saúde e, especificamente, no caso da "depressão", é essencial amparar-se nas referências médicas reconhecidas. O Manual Diagnóstico e Estatístico de Transtornos Mentais - DSM-V [American Psychiatric Association 2023] categoriza o grau de gravidade da depressão com base na ocorrência de um conjunto de sintomas. Por sua vez, o Inventário de Depressão de Beck (*Beck's Depression Inventory - BDI-II*) [Gorenstein and Andrade 1998] é um instrumento de avaliação que, respaldado na descrição dos sintomas depressivos abordados pelo DSM-V, tem sido utilizado na psicologia clínica para ajudar a mensurar o nível de severidade da depressão. O BDI-II, aplicado sob a forma de um questionário com 21 questões associadas a aspectos emocionais e comportamentais[2] de um indivíduo, produz uma pontuação que permite quantificar a intensidade dos sintomas depressivos em quatro graus de severidade: ausente, leve, moderada ou grave. Percebe-se que trabalhos que usam LLMs ajustados para a tarefa de classificação de severidade da depressão [Ji et al. 2022, Poświata and Perełkiewicz 2022] não contemplam todos os níveis especificados tanto pelo DSM-V, quanto pelo BDI-II, especialmente quanto ao grau "leve" de depressão. Dessa forma, a inclusão do grau "leve" permite que classificadores sejam estimulados a aprender sobre o que caracteriza cada nível de depressão em consonância com

[1]https://www.reddit.com/

[2]e.g., tristeza, pessimismo

os manuais médicos de referência.

Nesse contexto, o DepreBERTBR constitui uma iniciativa voltada à construção de um modelo de linguagem pré-treinado com textos de postagens no idioma português brasileiro, coletadas a partir de *subreddits* notórios por dispor de conteúdos com teor depressivo [Herculano et al. 2024b]. O modelo foi inicialmente ajustado e avaliado para a tarefa de classificação, em nível de postagem, considerando três graus de severidade de depressão (ausente, moderada e grave), utilizando o conjunto de dados desenvolvido por [Sampath and Durairaj 2022]. Torna-se oportuno realizar novos experimentos a fim de avaliar o desempenho do DepreBERTBR com a inclusão da classe "leve". Nesse sentido, uma das principais limitações para executar novos experimentos decorre da escassez de conjuntos de dados no idioma português brasileiro anotados com as quatro classes.

O `DepSeverity`[3] [Naseem et al. 2022] é um conjunto de dados no idioma inglês anotado por especialistas em saúde mental que contempla as quatro classes previstas no DSM-V para caracterização do nível de depressão: ausente, leve, moderado ou grave. Apesar de apresentar uma distribuição de frequência de classes desbalanceada, o que pode dificultar a discriminação das classes por parte do modelo, diante da dificuldade de dispor de outras alternativas de conjuntos de dados, encoraja-se investigar a utilização do `DepSeverity` com o suporte de técnicas de pré-processamento de dados.

O presente trabalho realiza um ciclo de experimentos utilizando o modelo DepreBERTBR e outros LLMs existentes na literatura, pré-treinados no idioma português brasileiro ou multilíngue, com o intuito de responder às seguintes questões de pesquisa:

- **QP1**: O DepreBERTBR é competitivo em relação a modelos de domínio geral em português ou multilíngue, ajustados para a tarefa de classificação de texto quanto à possível presença de um grau de depressão "ausente", "leve", "moderado" ou "grave"?
- **QP2**: Incrementar sinteticamente as instâncias do conjunto de dados `DepSeverity` melhora os resultados de predição dos LLMs envolvidos na comparação?

Para responder às questões de pesquisa anunciadas, este trabalho realiza uma avaliação comparativa entre os seguintes modelos pré-treinados com conteúdo no idioma português brasileiro: (i) DepreBERTBR [Herculano et al. 2024b], pré-treinado majoritariamente com conteúdo textual no domínio da depressão; (ii) BERTimbau [Souza et al. 2020] e BERTabaporu [Costa et al. 2023], ambos pré-treinados com conteúdo textual de domínio geral; e (iii) LaBSE [Feng et al. 2022] e mBERT [Devlin et al. 2019], modelos multilíngues pré-treinados com conteúdo textual em diferentes idiomas, além do português.

O artigo está organizado da seguinte forma: a Seção 2 provê um embasamento teórico; a Seção 3 descreve alguns trabalhos relacionados; a Seção 4 apresenta a metodologia de pesquisa empregada; a Seção 5 discute os resultados obtidos na avaliação experimental; e a Seção 6 tece algumas considerações e indica trabalhos futuros.

[3]https://github.com/usmaann/Depression_Severity_Dataset

2. Fundamentação Teórica

A evolução dos modelos de linguagem tem sido decisiva para o avanço de aplicações e técnicas baseadas em PLN. Esses modelos, de natureza probabilística ou neural, representam texto em formato numérico, de maneira que consigam capturar o contexto semântico e as relações entre as palavras [Caseli and Nunes 2024]. Modelos como o *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al. 2019] e o *Generative Pre-trained Transformer* (GPT) [OpenAI 2023] impulsonaram avanços significativos.

Neste trabalho, o modelo BERT, embasado na arquitetura *Transformer*, é utilizado para os estudos e comparações. Ele foi projetado para ser usado em duas etapas: pré-treinamento e ajuste fino. Na primeira etapa, o modelo não supervisionado é pré-treinado com base em um grande *corpus*, no qual aprende os padrões e nuances da linguagem usando dois objetivos de treinamento, a saber: *Masked Language Modeling* (Masked LM) e *Next Sentence Prediction* (NSP) [Feijó and Moreira 2020]. O Masked LM, por exemplo, foi usado para treinar o DepreBERTBR e consiste em mascarar aleatoriamente cerca de 15% dos tokens de entrada para, em seguida, realizar a previsão desses tokens. No ajuste fino, os parâmetros obtidos no pré-treinamento são reajustados conforme a tarefa alvo. No caso do DepreBERTBR, o ajuste fino foi realizado para a tarefa de classificação de texto em conjuntos de dados anotados para três [Sampath and Durairaj 2022] e quatro classes [Naseem et al. 2022] de nível de depressão.

Modelos de linguagem são muito influenciados pelo vocabulário do idioma e do domínio em que são treinados. Esse vocabulário pode ser monolíngue ou multilíngue, o que pode impactar a eficácia do modelo conforme regras sintáticas, semânticas e aspectos culturais [Caseli and Nunes 2024]. Assim, o uso de LLMs demanda a escolha entre modelos multilíngues pré-treinados com dados de diversos idiomas ou modelos específicos para um dado idioma e contexto. Como exemplo do primeiro caso, tem-se o *Multilingual BERT* (mBERT), treinado com textos da Wikipedia de 104 idiomas. Como modelos específicos no idioma português brasileiro, têm-se o DepreBERTBR e o BERTimbau.

No domínio específico da depressão no idioma português brasileiro, alguns trabalhos têm investido na criação de conjuntos de dados textuais, a exemplo do SetembroBR [Santos et al. 2023]. No entanto, esse conjunto de dados não foi concebido para fins de classificação de texto quanto ao grau de depressão. Em contrapartida, o `DepSeverity` [Naseem et al. 2022] surge como uma das poucas opções com anotações nas quatro classes de nível de depressão (ausente, leve, moderado e grave), mas no idioma inglês. O `DepSeverity` contempla a extração e filtragem de postagens coletadas do dataset CLEF eRisk[4], com o intuito de oferecer um suporte para a detecção de sinais de depressão em textos. Durante sua criação, foram empregados critérios de anotação alinhados ao BDI-II. Contudo, constata-se em sua distribuição um visível desbalanceamento de classes.

Para contornar esse problema, alguns estudos têm recorrido a técnicas de geração de dados sintéticos. O *Data Augmentation* [Guo et al. 2023] é uma técnica que vem sendo empregada para criar dados artificiais semelhantes a exemplos existentes no conjunto de dados original, seja utilizando LLMs como entidade geradora ou com a aplicação de regras específicas. Essa abordagem pode contribuir para a melhoria do desempenho dos modelos treinados a partir de conjuntos de dados desbalanceados [Li et al. 2023].

[4]https://early.irlab.org/

3. Trabalhos Relacionados

A literatura apresenta estudos que efetuaram comparações entre modelos de linguagem de domínio geral e específico na detecção de indícios de depressão, reservadas suas particularidades e idioma utilizado para seu pré-treinamento.

O trabalho de [Costa et al. 2023] desenvolveu o BERTabaporu, um modelo baseado na arquitetura BERT e pré-treinado com textos em português brasileiro abrangendo conteúdo sobre política, saúde mental e Covid-19. Seu desempenho foi comparado ao do BERTimbau [Souza et al. 2020] (ver Seção 4), em três tarefas preditivas: (a) postura crítica em relação a um tópico específico, (b) alinhamento político, e (c) estado de saúde mental. Na tarefa de saúde mental, a classificação foi realizada em nível de usuário em relação à ansiedade e depressão, com base em autodeclarações de diagnóstico feitas por usuários no Twitter/X. Os resultados apontaram que o BERTabaporu superou o BERTimbau em todas as métricas avaliadas (precisão, revocação e F1-*score*), destacando a importância da adaptação de modelos pré-treinados para contextos sensíveis, como a saúde mental.

[Silva et al. 2024] investigaram a eficácia da adaptação de modelos de linguagem em português brasileiro para a tarefa de classificação de texto no domínio governamental. Os autores ajustaram modelos linguísticos para aprimorar a compreensão das complexidades da linguagem jurídica e administrativa. Foram utilizados os modelos BERTimbau e LaBSE, em duas tarefas: (a) classificação de documentos com os conjuntos de dados LiPSet e SVic, e (b) classificação de itens, com os conjuntos de dados ProdServ e NaPEx. Em termos de avaliação, os modelos apresentaram um desempenho similar com base na métrica F1-Macro, com uma leve vantagem para o BERTimbau. Os autores ressaltam que fatores tais como o vocabulário técnico, jargões, relevância dos dados e o tamanho do conjunto de dados afetam o desempenho dos modelos em contextos específicos, sugerindo que as descobertas têm potencial para auxiliar na implementação de soluções automatizadas em setores públicos.

No contexto de análise de dados de redes sociais, [Skianis et al. 2024] examinaram a eficácia de modelos na previsão de grau de depressão ausente, leve, moderado e grave, utilizando dados em inglês. O estudo buscou investigar se o modelo conseguiria classificar indícios de depressão e ansiedade em dados originalmente em inglês e se manteriam sua eficácia com o respectivo conteúdo traduzido para outro idioma. Para isso, foi aplicada a técnica de Engenharia de *Prompts zero-shot* ao modelo GPT 3.5-turbo para realizar classificação a partir do conjunto de dados `DepSeverity`. Apesar da alta precisão alcançada pelo modelo (0,98), os F1-scores foram baixos (média de 0,17) mesmo no idioma original. Essa discrepância sugere que o modelo falha na cobertura geral das classes, tanto no idioma original quanto no traduzido. Os autores atribuem essa limitação ao desbalanceamento das classes no `DepSeverity` e à possibilidade de perda de nuances culturais e linguísticas ocorridas no processo de tradução automática.

Por fim, [Kang et al. 2024] exploraram a utilização de LLMs na geração de dados sintéticos a partir do conjunto de dados DAIC-WOZ, visando melhorar o desempenho preditivo para indicação de nível de depressão (ausente, presente). Utilizando o modelo Llama 3.2, os autores desenvolveram um *pipeline* composto por duas etapas: a geração de sinopses, i.e., resumos das transcrições originais de entrevistas clínicas, e a criação de variantes sintéticas dessas sinopses, ajustadas para diferentes níveis de severidade da de-

pressão. Devido à existência de um significativo desequilíbrio na distribuição das classes, a maioria dos pacientes foi classificada como não depressivos pelo modelo. Para mitigar esse problema, técnicas de *oversampling* foram empregadas durante a geração de dados sintéticos. Os resultados demonstraram que os modelos treinados com a inclusão de dados sintéticos superaram os treinados com dados reais.

O presente trabalho apresenta uma avaliação experimental com o intuito de examinar o comportamento de LLMs monolíngues (português brasileiro) e multilíngues ajustados para a tarefa específica de classificação de texto, considerando os quatro níveis citados de severidade de depressão. A pesquisa busca contribuir para o avanço de investigações e/ou abordagens relacionadas à detecção de sinais depressivos (em seu grau de intensidade) a partir de textos.

4. Metodologia

Esta seção descreve a metodologia de pesquisa empregada neste trabalho. Para tal, aborda os conjuntos de dados e modelos de linguagem utilizados para realizar a avaliação comparativa de desempenho quando ajustados para a tarefa de classificação de texto. As classes-alvo, amparadas nos critérios médicos, são aquelas definidas no BDI-II: ausente, leve, moderada ou grave.

4.1. Conjuntos de dados

Para avaliar o desempenho dos modelos quanto à tarefa de classificação, considerando o grau de severidade de depressão que pode estar embutido em relatos/textos de treinamento e teste, foi escolhido o conjunto de dados `DepSeverity` (Seção 1). Originalmente em inglês, o `DepSeverity` passou por um processo de tradução para o idioma português brasileiro utilizando a *Application Programming Interface* (API) do Google *Translate*. A Tabela 1 resume as estatísticas básicas do `DepSeverity` e sua distribuição de classes, em seu estado primário e após a inclusão de dados sintéticos. O conjunto de dados, originalmente, possui um total de 3.553 instâncias. Nota-se a existência de um desbalanceamento acentuado na proporção de distribuição das classes, particularmente nas classes de interesse que sinalizam maior intensidade da depressão.

Tabela 1. Visão geral do dataset DepSeverity Original e Aumentado.

Dados originais		**Dados originais + sintéticos**	
Classe	**Distribuição**	**Classe**	**Distribuição**
Ausente	2.587 (73%)	Ausente	2.069 (31%)
Leve	290 (8%)	Leve	1351 (21%)
Moderada	394 (11%)	Moderada	1861 (29%)
Grave	282 (8%)	Grave	1226 (19%)
Total	**3.553 (100%)**	**Total**	**6.507 (100%)**

Haja vista o visível desbalanceamento de classes que pode comprometer o desempenho dos modelos em predizer corretamente as classes leve, moderada e grave (além de que, durante a divisão dos dados em treino, validação e teste, o número de instâncias das classes minoritárias pode se tornar insuficiente), buscou-se mitigar o problema exposto com a produção de instâncias sintéticas das classes minoritárias. [Naseem et al. 2022]

mencionam que a complexidade de classificar níveis de depressão ordinalmente (ausente <leve <moderada <grave) pode tornar difícil a aplicação de técnicas de balanceamento de dados tradicionais, visto que geralmente essas abordagens hierárquicas assumem uma distribuição mais uniforme entre as classes. Com isso, foi aplicada uma técnica de *Data Augmentation* (Seção 2).

O acréscimo de novas instâncias ao conjunto de dados original `DepSeverity` foi realizado utilizando-se o modelo de linguagem Gemma 2[5] do Google, devido à sua alta capacidade na geração de textos e disponibilidade no Huggingface[6], de tal forma a reescrever as instâncias apenas dos dados do conjunto de treino e validação, excluindo-se a partição de teste. Essa exclusão teve como principal objetivo preservar os textos do conjunto de teste inalterados, uma vez que representariam dados futuros em que o modelo deveria colocar em prática o seu poder de generalização. Empregando-se a técnica de *few-shot* em um prompt adaptado das instruções usadas por [Li et al. 2023], foram incluídos exemplos de instâncias das classes minoritárias de tal modo a orientar o modelo Gemma 2 sobre como as novas instâncias deveriam ser geradas. Sendo assim, foram geradas e adicionadas cinco novas instâncias para cada uma das instâncias originais das classes minoritárias. Com essa proporção, vislumbrou-se uma distribuição menos desequilibrada das classes do problema. Dessa forma, os conjuntos de treino e validação, considerando os dados originais do `DepSeverity` em conjunto com as instâncias sintéticas, somaram 6.507 instâncias (Tabela 1).

Para avaliação da similaridade dos textos produzidos sinteticamente em relação ao original, optou-se por usar a métrica BERTScore. Esta métrica utiliza *embeddings* de linguagem gerados pelo modelo BERT para calcular a similaridade entre as sentenças geradas pelo modelo e aquelas utilizadas como referência [Zhang et al. 2019] e, com base nisso, produz o *score* indicando o grau de correspondência semântica entre os textos, considerando toda a sentença em questão [Xia et al. 2015]. A geração das instâncias sintéticas apresentou um valor de BERTScore = 0,91 em relação às instâncias originais, o que demonstra alta similaridade em termos de contexto e semântica.

4.2. Modelos pré-treinados

Os modelos escolhidos para a avaliação experimental e apresentados nesta seção são baseados na arquitetura BERT [Devlin et al. 2019]. Todos possuem as características de terem sido pré-treinados com textos específicos no domínio da depressão ou com textos de domínio geral no idioma português brasileiro. Alguns modelos, também treinados com textos de domínio geral, foram selecionados por sua capacidade de processar dados multilíngue, tendo o português como um dos idiomas suportados. Cabe destacar que, até o período da escrita deste trabalho, não foram encontrados na literatura outros LLMs pré-treinados no idioma português brasileiro especializados no domínio único da depressão. Por isso, justifica-se a decisão de examinar o comportamento de outros LLMs disponíveis também no idioma português brasileiro (mas de domínio geral) quanto à tarefa de classificação de textos com indícios de depressão. Devido a questões de limitação de recursos computacionais, haja vista que treinar modelos de linguagem muitas vezes requer máquinas robustas, hospedadas em servidores na nuvem e com alto custo financeiro, todos os modelos foram instanciados na versão *Base* do BERT. Além disso, até o

[5]https://huggingface.co/docs/transformers/main/en/model_doc/gemma2

[6]https://huggingface.co/

momento de desfecho dos experimentos, alia-se o fato de que os modelos mBERT e o LaBSE só estavam disponíveis na versão *Base*.

DepreBERTBR [Herculano et al. 2024b]: Modelo de linguagem pré-treinado com o conjunto de dados DepreRedditBR[7]. Este conjunto de dados possui 509.675 postagens no idioma português brasileiro, com teor depressivo, não anotadas, coletadas a partir de publicações na rede social Reddit [Herculano et al. 2024a]. As postagens foram publicadas entre janeiro de 2018 e outubro de 2023.

BERTimbau [Souza et al. 2020]: Modelo de linguagem pré-treinado com o conjunto de dados brWaC (Brazilian Web as Corpus)[Wagner Filho et al. 2018], um conjunto de dados textuais constituído de 145 milhões de sentenças extraídas de páginas web brasileiras na internet.

BERTabaporu [Costa et al. 2023]: Modelo de linguagem pré-treinado utilizando textos no idioma português brasileiro extraído de postagens coletadas da rede social Twitter (atualmente denominado X[8]). O conjunto de dados utilizado no pré-treino compreende 238 milhões de sentenças de domínio geral, contemplando postagens diversificadas em temáticas sobre política, Covid-19 e saúde mental.

LaBSE [Feng et al. 2022]: *Language-agnostic BERT Sentence Embedding*, o LaBSE é um modelo multilingue pré-treinado com textos escritos em vários idiomas, de domínio geral, inclusive no idioma português. O seu treinamento utilizou uma base de dados constituída por 17 bilhões de sentenças oriundas da Wikipédia[9] e CommonCrawl[10] para a tarefa de *Masked Language Modeling*. No que diz respeito à tarefa específica de tradução (*Translation Language Modeling*), o LaBSE utilizou um conjunto de dados com 6 bilhões de pares de sentenças traduzidas.

mBERT [Devlin et al. 2019]: O modelo de linguagem *multilingual* BERT é uma versão original do BERT pré-treinado com dados textuais de 104 idiomas, incluindo o português brasileiro. Os dados textuais de pré-treinamento foram coletados a partir de artigos da Wikipedia em diferentes idiomas. O tamanho do conjunto de dados de pré-treinamento do mBERT não foi informado pelos autores.

A Tabela 2 ilustra uma visão geral das características mencionadas dos modelos descritos e usados na avaliação alvo deste trabalho. Nota-se que há uma grande diferença no tamanho dos conjuntos de dados usados no pré-treinamento quando se compara o modelo DepreBERTBR e os demais. Observa-se também que apenas o BERTabaporu, mesmo sendo de domínio geral, utilizou dados relacionados à saúde mental em seu pré-treinamento. Os modelos utilizados como *baseline* foram selecionados por terem sido treinados também em português brasileiro, sendo, portanto, o DepreBERTBR o único especializado no domínio da depressão neste idioma.

4.3. Protocolo de avaliação

Para avaliação dos LLMs supracitados, todos foram ajustados para a tarefa de classificação considerando quatro classes de nível de intensidade de depressão (ausente,

[7]https://zenodo.org/records/12761179
[8]https://www.x.com/
[9]https://www.wikipedia.org/
[10]https://commoncrawl.org/

Tabela 2. Visão geral dos modelos de linguagem comparados

Modelo	Domínio	Origem	Tamanho do corpus	Idioma
DepreBERTBR	Depressão	Reddit	509.675 mil	Português
BERTimbau	Geral	Web	145 milhões	Português
BERTabaporu	Geral + Saúde mental	Twitter	238 milhões	Português
LaBSE	Geral	Wikipedia + CommonCrawl	23 bilhões	Multilíngue
mBERT	Geral	Wikipedia	Não informado	Multilíngue

leve, moderada, grave), seguindo as referências médicas e, em particular, o BDI-II. O `DepSeverity` foi particionado utilizando-se a técnica de validação *hold-out* estratificada, com 80% dos dados usados para treinamento e validação (realização do ajuste fino) e os 20% restantes utilizados para teste. Os modelos avaliados foram configurados para treinamento em duas épocas, utilizando o otimizador AdamW com uma função de ativação *softmax* na camada de saída e taxa de aprendizado igual a 5e-5, configurações estas adotadas como padrão do modelo BERT no *Hugging face*. Os experimentos foram realizados utilizando como ambiente computacional a plataforma Google Colaboratory Pro+ e uma GPU NVIDIA Ampere A100 Tensor Core. Para avaliar o desempenho dos modelos na tarefa de classificação, foram calculadas as medidas de precisão, revocação e *F1-score* para cada classe, de todos os modelos comparados. Em seguida, foram calculadas as médias dos resultados de todas as métricas citadas.

5. Resultados e discussão

Os experimentos para avaliação dos modelos de linguagem foram divididos em dois cenários. O primeiro utilizou a distribuição de dados original do conjunto de dados `DepSeverity`. O segundo cenário utilizou os dados originais do `DepSeverity` com o acréscimo das instâncias das classes minoritárias geradas sinteticamente, conforme descrito na Seção 4. A Tabela 3 apresenta uma visão geral dos resultados obtidos após o ajuste fino dos modelos para a tarefa de classificação nas quatro classes de nível de depressão. São reportadas, para cada modelo avaliado, as métricas de precisão (P), revocação (R) e *F1-score* (F1) por classe, considerando os dados íntegros do `DepSeverity`.

Tabela 3. Avaliação dos modelos com o conjunto de dados Depseverity original

	DepreBERTBR			**BERTimbau**			**BERTabaporu**			**LaBSE**			**mBERT**		
Classes	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Ausente	0,83	0,91	0,87	0,85	0,87	0,86	0,84	0,89	0,87	0,83	0,87	0,85	0,83	0,87	0,85
Leve	0,46	0,32	0,38	0,36	0,34	0,35	0,31	0,34	0,32	0,42	0,41	0,42	0,37	0,15	0,21
Moderada	0,50	0,29	0,36	0,47	0,31	0,37	0,54	0,30	0,39	0,41	0,37	0,39	0,30	0,31	0,31
Grave	0,32	0,32	0,32	0,29	0,39	0,33	0,33	0,32	0,33	0,36	0,25	0,29	0,31	0,35	0,33
AVG	**0,52**	**0,46**	**0,48**	0,49	0,48	0,48	0,51	0,46	0,48	**0,51**	**0,47**	**0,49**	0,45	0,42	0,43

Os resultados demonstram que todos os modelos apresentaram um bom desempenho na predição correta da classe "ausente", como era esperado, com F1 = 0,87. Porém, constata-se que os modelos apresentaram um $F1$ insatisfatório para as classes "leve", "moderada"e "grave", conforme valores apresentados na Tabela 3. A principal suposição para esta limitação está relacionada à baixa representatividade dessas classes no conjunto de dados, em quantidade insuficiente para que o modelo conseguisse perceber padrões consistentes inerentes às características de cada classe do problema. Alia-se a isso, ainda,

o fato de que a divisão de dados estratificada pela técnica *hold-out* reduziu a ocorrência destas instâncias minoritárias nos conjuntos de treino e teste, prejudicando a capacidade de generalização do modelo.

A matriz de confusão ilustrada na Figura 1 corrobora com a suspeita de que os classificadores de todos os modelos estão enviesando a predição para a classe majoritária, haja vista o evidente desbalanceamento de classes. Nota-se que os resultados falsos negativos das classes leve, moderada e grave são dominantes para classificações equivocadas. O resultado geral indica uma grande dificuldade dos modelos em prever corretamente as classes indicativas de nível de depressão.

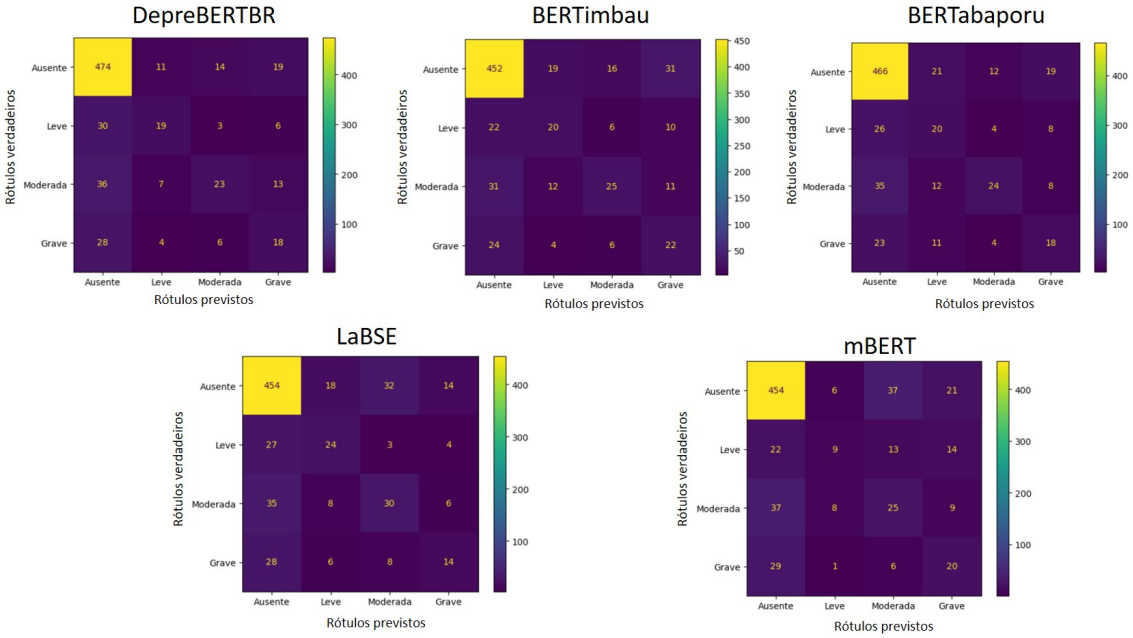

Figura 1. Matriz de confusão com dados do Depseverity Original

Em cenários desbalanceados, a literatura recomenda a aplicação de técnicas de reamostragem. Uma delas, a *oversampling* fundamenta-se basicamente no incremento de instâncias da classe minoritária. Ao aplicarmos a técnica de *Data Augmentation* para essa finalidade de mitigação do desbalanceamento das classes leve, moderada e grave no conjunto de dados `Depseverity`, os modelos de linguagem foram submetidos a um novo ajuste fino.

A Tabela 4 demonstra que, mesmo com a tentativa de balanceamento de classes do `DepSeverity`, todos os modelos não demonstraram melhoria de desempenho na classificação das classes minoritárias. Os equívocos inerentes aos falsos negativos seguem o mesmo padrão observado na Figura 1.

Alguns pressupostos podem ser levantados para explicar esses resultados. Haja vista que o `DepSeverity` é disponibilizado originalmente no idioma inglês, pode ter ocorrido ruído no processo de tradução, causando a perda do contexto e/ou sentido dos textos originais. A inexistência de conjuntos de dados anotados nas quatro classes de depressão (ausente, leve, moderada e grave) originalmente no idioma português brasileiro implica na utilização de conjuntos de dados traduzidos de outros idiomas como o inglês,

sendo necessário utilizar ferramentas de tradução automáticas. Realizar essa tradução de forma manual e precisa demanda muito tempo e esforço humano. Além disso, implicaria em custos adicionais e no envolvimento de especialistas que avaliassem não só a tradução, mas a correlação entre os textos originais e os textos traduzidos, adequados ao idioma de destino. Outra teoria plausível pode ser atribuída à qualidade da anotação das instâncias feitas pelos especialistas atuantes no contexto do `Depseverity`, uma vez que os autores desse conjunto de dados indicam que os relatos analisados podem conter vieses em decorrência da avaliação feita pelos especialistas anotadores [Naseem et al. 2022].

Diante disso, é possível que instâncias com características equivalentes tenham sido rotuladas com classes distintas. Esses ruídos nas anotações e os textos selecionados para a criação do `Depseverity` podem ter dificultado o aprendizado dos modelos, não permitindo que algumas nuances psicológicas sensíveis ao domínio da depressão fossem compreendidas durante o ajuste fino. Isso também pode ter produzido impacto na geração das instâncias sintéticas que, apesar de terem sido geradas com uma alta similaridade semântica, podem ter absorvido possíveis ruídos, causados tanto pela tradução, quanto pela anotação dos textos. Dessa forma, os resultados sugerem que o `Depseverity` impõe algumas limitações para certos tipos de tarefas (e.g., classificação no nível de severidade de depressão), que devem ser investigadas e resolvidas caso a caso com a exploração de novas técnicas e/ou abordagens para mitigar os problemas ocasionados pelo seu desbalanceamento natural de classes.

Tabela 4. Avaliação dos modelos com o conjunto de dados Depseverity aumentado

	DepreBERTBR			**BERTimbau**			**BERTabaporu**			**LaBSE**			**mBERT**		
Classes	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Ausente	0,82	0,85	0,84	0,84	0,84	0,84	0,83	0,88	0,85	0,83	0,84	0,84	0,81	0,89	0,85
Leve	0,28	0,18	0,22	0,29	0,34	0,31	0,33	0,32	0,33	0,22	0,17	0,19	0,25	0,24	0,25
Moderada	0,35	0,31	0,33	0,40	0,34	0,37	0,45	0,34	0,39	0,36	0,43	0,39	0,29	0,22	0,25
Grave	0,29	0,33	0,31	0,33	0,33	0,33	0,33	0,28	0,30	0,35	0,30	0,32	0,41	0,21	0,28
AVG	0,43	0,42	0,43	0,47	0,46	0,46	**0,49**	**0,45**	**0,47**	0,44	0,43	0,43	0,41	0,39	0,41

Considerados os resultados reportados pelos experimentos, as questões de pesquisa formuladas neste trabalho são respondidas como segue:

- **QP1**: embora o DepreBERTBR tenha sido pré-treinado com um conjunto de dados indiscutivelmente inferior em relação aos que foram utilizados para pré-treinar os modelos comparados, este se mostrou competitivo em termos de desempenho (Tabela 3), demonstrando atuação equiparável em termos de média de *F1* (predição de todas as classes). O mesmo desempenho também é observado no segundo cenário de experimentação, onde são adicionados dados sintéticos ao `DepSeverity` (Tabela 4). É importante salientar que o tamanho do conjunto de dados utilizado no pré-treinamento pode impactar significativamente o desempenho do modelo, pois, geralmente, quanto maior a quantidade de dados disponíveis, maior o conhecimento adquirido pelo modelo pré-treinado. Consequentemente, quanto mais dados um modelo aprende, mais conhecimento transmitido através do ajuste fino para tarefas como a classificação.
- **QP2**: o aumento do `DepSeverity` com dados sintéticos não melhorou os resultados de predição, tanto para o DepreBERTBR quanto para os demais modelos comparados. Isso porque o `DepSeverity` pode apresentar limitações

em seu conteúdo original, onde sentenças podem conter ambiguidades que causam divergências entre o contexto e o rótulo anotado. Além disso, a tradução do `DepSeverity`, mencionada anteriormente, pode ter produzido distorções semânticas no sentido real das instâncias, refletindo no aprendizado dos modelos e provocando um desempenho abaixo do esperado no tocante à comparação tanto no cenário com dados originais, quanto na própria tarefa de classificação.

6. Considerações e trabalhos futuros

Este trabalho apresentou uma avaliação experimental comparativa entre o modelo de linguagem pré-treinado DepreBERTBR e os modelos pré-treinados de domínio geral em português brasileiro BERTimbau e BERTabaporu. Adicionalmente, comparou-se o DepreBERTBR com os modelos multilíngue LaBSE e mBERT, ambos também de domínio geral.

Os resultados demonstraram que os modelos comparados apresentaram dificuldades em classificar corretamente as classes referentes aos graus de depressão "leve", "moderado" e "grave" após ajuste fino utilizando-se o conjunto de dados `DepSeverity`, mesmo no cenário em que foram acrescentados exemplos sintéticos na tentativa de alcançar um cenário de distribuição de classes mais balanceado. Em termos das métricas precisão, revocação e F1, certos modelos apresentam melhor discriminação de uma classe em relação às demais. Apesar do resultado das avaliações não apresentar valores discrepantes para os dois cenários experimentais investigados, observa-se que o treinamento dos modelos sem o acréscimo de exemplos sintéticos apresentou melhorias sutis na avaliação do conjunto de teste. O DepreBERTBR mostra-se, portanto, uma opção viável para o problema investigado, haja vista que seu pré-treinamento com dados específicos do domínio da depressão oferece uma opção de modelo com menor custo de produção (tempo e recursos de hardware), quando comparado aos demais modelos com resultados equivalentes.

Uma limitação apontada neste trabalho diz respeito ao conjunto de dados `DepSeverity`. O processo de tradução automática de texto pode ter introduzido ruídos nas instâncias, provocando perdas do sentido original dos textos e, consequentemente, influenciado na geração dos dados sintéticos. Em relação à tentativa de balanceamento de classes aplicada ao `DepSeverity` em seu conjunto de treinamento e validação, os resultados não confirmaram a melhoria esperada, sugerindo a busca por conjuntos de dados alternativos ou aplicação de outras técnicas de mitigação de desbalanceamento.

Como trabalhos futuros, objetiva-se investir na anotação do conjunto de dados DepreRedditBR conforme as quatro classes do BDI-II e realizar uma nova avaliação comparativa entre os modelos utilizando, desta vez, o DepreRedditBR anotado.

Agradecimentos. Este trabalho foi desenvolvido com apoio da FAPESQ-PB.

Referências

American Psychiatric Association (2023). *Manual diagnóstico e estatístico de transtornos mentais: DSM-5-TR*. Artmed, Porto Alegre, 1 edition. Tradução da obra original: *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*, 2022.

Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 3 edition.

Costa, P. B., Pavan, M. C., Santos, W. R., Silva, S. C., and Paraboni, I. (2023). Bertabaporu: assessing a genre-specific language model for portuguese nlp. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 217–223.

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 American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. NAACL.

Feijó, D. D. V. and Moreira, V. P. (2020). Mono vs multilingual transformer-based models: a comparison across several language tasks. *arXiv*, 2007.09757v1.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT sentence embedding. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Gorenstein, C. and Andrade, L. (1998). Inventário de depressão de beck: propriedades psicométricas da versão em português. *Rev psiq clin*, 25(5):245–50.

Guo, Z., Wang, P., Wang, Y., and Yu, S. (2023). Dr. llama: Improving small language models on pubmedqa via generative data augmentation. *ArXiv*, abs/2305.07804.

Herculano, A., de Paula, T.-H., Fernandes, D., and Rego, A. (2024a). Deprereddittbr: Um conjunto de dados textuais com postagens depressivas no idioma português brasileiro. In *Anais do VI Dataset Showcase Workshop*, pages 77–90, Porto Alegre, RS, Brasil. SBC.

Herculano, A., Souza, D., and Rego, A. (2024b). Deprebertbr: Um modelo de linguagem pré-treinado para o domínio da depressão no idioma português brasileiro. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 181–194, Porto Alegre, RS, Brasil. SBC.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., and Cambria, E. (2022). MentalBERT: Publicly available pretrained language models for mental healthcare. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Kang, A., Chen, J. Y., Lee-Youngzie, Z., and Fu, S. (2024). Synthetic data generation with llm for improved depression prediction. *ArXiv*, abs/2411.17672.

Li, Z., Zhu, H., Lu, Z., and Yin, M. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Naseem, U., Dunn, A. G., Kim, J., and Khushi, M. (2022). Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, pages 2563–2572.

OMS (2023). Organização mundial de saúde (oms): Desordem depressiva (depressão). https://www.who.int/news-room/fact-sheets/detail/depression. Último Acesso 23 de Abr 2025.

OpenAI (2023). Gpt-4 technical report. Technical report, OpenAI. Accessed on April 2, 2025.

Pavlopoulos, A., Rachiotis, T., and Maglogiannis, I. (2024). An overview of tools and technologies for anxiety and depression management using ai. *Applied Sciences*, 14(19).

Poświata, R. and Perełkiewicz, M. (2022). Opi@ lt-edi-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282.

Sampath, K. and Durairaj, T. (2022). Data set creation and empirical analysis for detecting signs of depression from social media postings. In *International Conference on Computational Intelligence in Data Science*, pages 136–151. Springer.

Santos, W. R. d., de Oliveira, R. L., and Paraboni, I. (2023). Setembrobr: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, pages 1–28.

Silva, M., Oliveira, G., Costa, L., and Pappa, G. (2024). Evaluating domain-adapted language models for governmental text classification tasks in portuguese. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 247–259, Porto Alegre, RS, Brasil. SBC.

Skianis, K., Doğruöz, A. S., and Pavlopoulos, J. (2024). Leveraging LLMs for translating and classifying mental health data. In Sälevä, J. and Owodunni, A., editors, *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 236–241, Miami, Florida, USA. Association for Computational Linguistics.

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.

Uban, A.-S., Chulvi, B., and Rosso, P. (2021). An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: a new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Xia, P., Zhang, L., and Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences*, 307:39–52.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.