DepreBERTBR: Um Modelo de Linguagem Pré-treinado para o Domínio da Depressão no Idioma Português Brasileiro

Ayrton Douglas Rodrigues Herculano[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@academico.ifpb.edu.br, {damires,alex}@ifpb.edu.br

Abstract. *Depression has been a growing concern in modern society and, according to the World Health Organization, this disease may become the most common by 2030. Previously restricted to medical' offices, depressive feelings have also been shared on social networks such as Reddit. In this light, this work presents an approach for classifying social media posts with depressive signs. It is based on the creation of a corpus and a pre-trained language model called DepreBERTBR, considering the Brazilian Portuguese language. The DepreBERTBR has been tuned for such classification task according to three possible levels of depression: absent, moderate or severe. The results demonstrate that DepreBERTBR is competitive w.r.t. other portuguese language models.*

Resumo. *A depressão tem sido alvo de preocupação na sociedade moderna e, conforme a OMS, pode se tornar a doença mais comum até 2030. Antes restritos aos consultórios, sentimentos com teor depressivo têm sido compartilhados em redes como a Reddit. Neste cenário, este trabalho propõe uma abordagem para classificação de postagens de redes sociais com sinais de depressão, que se apoia na construção de um corpus e de um modelo de linguagem pré-treinado chamado DepreBERTBR, considerando o idioma português brasileiro. O DepreBERTBR foi ajustado para a tarefa citada conforme três graus de depressão: ausente, moderada ou grave. Os resultados demonstram que o DepreBERTBR é competitivo com respeito a outros modelos de linguagem em português.*

1. Introdução

A depressão é um transtorno mental que afeta a natureza emocional, psicológica e física de um indivíduo. Alguns sinais que indicam um padrão de suspeita de desenvolvimento da depressão incluem [OMS 2023]: humor deprimido, perda de prazer ou interesse em realizar atividades, sentimento de culpa e baixa autoestima. Fatores de risco como histórico familiar, obesidade e luto podem contribuir com o desenvolvimento da doença. É notório que a depressão já é uma doença do cotidiano da população, sendo até mesmo intitulada como o Mal do Século pela Organização Mundial de Saúde (OMS). A própria OMS estima que, até o ano de 2030, a depressão será a doença mais comum em escala mundial.

A depressão pode se manifestar em pessoas de qualquer raça, sexo ou idade, minando sua capacidade de conduzir uma vida normal. Reconhecer pessoas com depressão não é uma tarefa fácil, haja vista que há o receio delas esconderem seus sintomas por medo de serem julgadas ou até mesmo fingirem que está tudo bem, apresentando-se

felizes para outras pessoas mas, internamente, se sentindo desanimadas. As situações mencionadas podem ser compreendidas como motivos para que pessoas afetadas por depressão dificultem o primeiro contato com um médico especialista para avaliação. O diagnóstico precoce pode ajudar significativamente no tratamento e cura da doença [American Psychiatric Association 2013].

O monitoramento da atividade de usuários em redes sociais pode auxiliar na descoberta de tendentes a um comportamento depressivo, pois as postagens compartilhadas publicamente podem ser ricas em emoções reveladoras [Cacheda et al. 2019]. Estudos mostram que há um crescente interesse no uso de redes sociais em busca de um bem-estar social e fisiológico, pois as redes se tornaram um espaço virtual popular para os mais diversos fins (e.g., lazer, opiniões). Considerando que o uso das redes sociais fomenta a possibilidade de socialização em um ambiente controlado, os indivíduos com depressão podem se sentir mais atraídos pelas interações nas redes sociais do que pelas interações presenciais. [Uban et al. 2021] ressaltam que usuários acometidos pela depressão se sentem mais à vontade para expor seus sentimentos quando interagem em espaços focados na discussão sobre o tema, seja em busca de apoio ou para se identificarem com outros usuários que estão passando pelo mesmo problema. A rede social Reddit[1] é um ambiente que corrobora com essa evidência, pois dispõe de subcomunidades para debates específicos, alguns deles relacionados ao domínio de transtornos mentais (ver Seção 4.1).

Por se tratar de uma área multidisciplinar, esforços oriundos das áreas de Psiquiatria e Psicologia associados ao uso de técnicas computacionais do campo do Processamento de Linguagem Natural (PLN) e de Análise de Sentimentos vêm sendo empregados com o intuito de reconhecer textos com teor depressivo. Para aumentar o desafio, a depressão pode ser classificada clinicamente em relação ao seu nível de severidade em [de Psiquiatria 2022]: ausente (sem depressão), leve, moderada ou grave. Recentemente, modelos de linguagem de grande porte (do inglês, *Large Language Model* - LLM) vêm sendo comumente explorados em problemas de análise de sentimentos, buscando identificar com maior precisão sentimentos embutidos em textos, particularmente em postagens de redes sociais [Costa et al. 2023]. Os LLMs são modelos de linguagem treinados a partir de um enorme volume de dados textuais com o objetivo de aprender padrões sobre como as palavras são usadas em sentenças de um determinado idioma. Exemplos de tarefas em que LLMs são aplicados incluem tradução de idiomas e sumarização de texto.

É possível encontrar pesquisas na literatura que empregam LLMs para detectar sinais de depressão em postagens de redes sociais [Ji et al. 2022, Poświata and Perełkiewicz 2022]. Normalmente, durante a construção do modelo, pode ser necessária a preparação de um *corpus* para treinamento do modelo. Neste cenário, a maioria dos trabalhos ainda é focado no idioma inglês, o que impossibilita seu reuso em problemas que precisam lidar com o reconhecimento e interpretação da linguagem humana no idioma português. Diante do exposto, duas Questões de Pesquisa (QP) norteiam o presente trabalho, voltado à análise de postagens no idioma português brasileiro no domínio da depressão:

- **QP1**: Como classificar postagens de redes sociais considerando diferentes níveis de severidade de depressão?

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

- **QP2**: Utilizar um modelo de linguagem pré-treinado no domínio específico da depressão pode ser útil para classificar o grau de severidade da depressão embutido em postagens de redes sociais?

Para responder às questões postas, este trabalho propõe uma abordagem para classificação de postagens de redes sociais com tendências depressivas. A presente abordagem se apoia na construção de um *corpus* e de um modelo de linguagem pré-treinado denominado DepreBERTBR, considerando o idioma português brasileiro. O DepreBERTBR foi ajustado para a tarefa de classificação de postagens conforme três graus de severidade de depressão: ausente, moderada ou grave. Os resultados obtidos, por meio de uma avaliação experimental, demonstram que o DepreBERTBR é competitivo com respeito a outros modelos de linguagem no referido idioma. Ressalta-se, adicionalmente, que o pré-treinamento de um LLM com um *corpus* específico do domínio da depressão é oportuno para alcançar resultados promissores em problemas de classificação do nível de severidade de depressão.

O restante do artigo está organizado da seguinte forma. A Seção 2 introduz conceitos pertinentes. A Seção 3 descreve alguns trabalhos relacionados. A Seção 4 apresenta a abordagem proposta, e a Seção 5 discute os resultados obtidos na avaliação experimental. Por fim, a Seção 6 tece algumas considerações e indica trabalhos futuros.

2. Fundamentação teórica

Em tempos recentes, modelos de aprendizado profundo têm sido amplamente empregados, inclusive conjuntamente com tarefas de PLN e análise de sentimentos. Alguns aspectos favoráveis a isso são [Oliveira et al. 2022]: (i) exigem pouca engenharia de features; e (ii) produzem representações vetoriais que capturam similaridades de unidades linguísticas facilitando, assim, o entendimento do conhecimento.

Neste cenário, outro conceito cada vez mais discutido e utilizado se refere aos mecanismos de atenção, um conjunto adicional de parâmetros para uma rede neural onde os itens mais relevantes da entrada recebem uma valoração maior no vetor de contexto [Caseli and Nunes 2023]. Aliado a isso, diante do aumento de dados e da evolução de recursos computacionais, cada vez mais pode-se considerar tarefas de aprendizado profundo que usufruam de etapas de pré-treinamento de modelos para transferência de aprendizado [Pan and Yang 2009]. No contexto atual deste paradigma, um modelo de aprendizado profundo pode ser pré-treinado como um LLM a partir do uso de grandes conjuntos de dados. Como ilustração, um LLM pode ser pré-treinado com base em um *corpus* com milhões de sentenças do domínio de saúde mental. O LLM pré-treinado é, então, adaptado a diferentes tarefas posteriores por meio da definição de parâmetros adicionais, ajustando-o a partir da tarefa alvo que pode ser, por exemplo, uma classificação de sentimentos binária ou multiclasse. Após aprender as características da linguagem como contexto, gramática e idioma, esse modelo base pode ser ajustado para realizar, por exemplo, uma classificação de sentimentos associados a ansiedade, depressão ou estresse, utilizando um conjunto de dados rotulados com essas classes em tamanho menor.

LLMs podem ser baseados em *Transformer*, uma arquitetura fundamentada em rede neural profunda, que utiliza uma estrutura codificador-decodificador com um mecanismo de auto-atenção para aprender a relação complexa entre palavras de um texto [Vaswani et al. 2017]. O componente codificador transforma o texto de entrada em

uma representação vetorial, enquanto que o decodificador converte a representação para um texto de saída. É possível adicionalmente não utilizar todos componentes existentes na sua arquitetura, e sim partes deles [Caseli and Nunes 2023]. Este trabalho é baseado no modelo BERT (*Bidirectional Encoder Representations for Transformers*) [Devlin et al. 2019]. O BERT é uma das instanciações mais utilizadas atualmente e considera apenas o componente codificador de um *transformer*. Ele foi treinado em duas versões: *Base* com 12 camadas de *transformers* e *Large* com 24 camadas.

Informações mais detalhadas sobre o *corpus* em inglês e pré-treinamento do BERT podem ser obtidas em [Devlin et al. 2019]. Em particular, o BERT utiliza o tokenizador WordPiece [Wu et al. 2016], o qual transforma o texto em uma sequência de tokens (palavras e/ou subpalavras) para construir um vocabulário de tokens únicos. O BERT foi implementado em duas etapas: Pré-treino e Ajuste fino (*Fine tuning*). Na etapa do pré-treino o modelo é treinado com dados textuais para aprender sobre o contexto utilizando duas tarefas não supervisionadas: *Masked Language Modeling* e *Next Sentence Prediction*. A primeira, utilizada neste trabalho, consiste em mascarar, de forma aleatória, um percentual dos tokens de entrada (em torno de 15%) e depois realizar a previsão desses tokens. Na etapa de Ajuste Fino, o modelo é iniciado com os valores dos parâmetros do pré-treino e, depois, os parâmetros são reajustados conforme uma tarefa específica.

3. Trabalhos Relacionados

Estudos têm sido realizados em busca de classificar postagens de usuários em redes sociais com respeito a possíveis indicativos de depressão. Diante da abundância de dados existentes em redes sociais, da possibilidade de coletá-los e da contínua evolução de técnicas associadas ao paradigma de pré-treinamento de LLMs [Liu et al. 2023], alguns trabalhos passaram a propor soluções de classificação de textos que fazem uso de LLMs pré-treinados em domínios gerais ou específicos como o da depressão. Um desafio no desenvolvimento dessas soluções diz respeito ao *corpus* a ser usado, tendo em vista que, dependendo do domínio em questão, haverá ou não algum que esteja disponível, principalmente em um idioma em particular. Em casos de disponibilidade, pode acontecer do *corpus* não dispor de um número de exemplos suficientes para realizar o pré-treinamento do modelo, o que implica na necessidade de sua construção. Particularmente, quando se trata do idioma "português brasileiro", ainda são escassos conjuntos de dados com conteúdo proveniente de postagens de redes sociais de cunho depressivo, assim como são poucos os trabalhos baseados em modelos pré-treinados para o domínio da depressão.

Considerando o idioma inglês e o domínio específico da depressão, o trabalho de [Ji et al. 2022] desenvolveu dois modelos de linguagem denominados MentalBERT e MentalRoBERTa. O primeiro modelo foi inicializado com o BERT-Base. O MentalRoBERTa, por sua vez, é um modelo inicializado com o RoBERTa-Base [Liu et al. 2019], modelo que elimina a tarefa de *Next Sentence Prediction* do processo de treinamento original do BERT. Ambos os modelos foram treinados com um *corpus* de 13 milhões de postagens relacionadas à saúde mental coletadas do Reddit. Após o pré-treinamento dos modelos, foi realizado o ajuste fino para a tarefa de classificação de textos, levando em conta transtornos mentais como a depressão, ansiedade, ideação suicida e estresse. Os dois modelos desenvolvidos foram avaliados com respeito a outros modelos existentes, o BERT e o RoBERTa, pré-treinados com um *corpus* geral, e o BioBERT pré-treinado no domínio biomédico [Lee et al. 2020]. Como resultado, o pré-treinamento dos modelos

com um *corpus* do domínio-alvo de saúde mental se mostrou mais útil em relação aos modelos pré-treinados com *corpus* do domínio biomédico e de domínio geral, haja vista que apresentou melhor desempenho em tarefas de classificação quanto à saúde mental.

O DepRoBERTa (RoBERTa para detecção de depressão) é um modelo de linguagem baseado no RoBERTa-large e pré-treinado em postagens depressivas extraídas do Reddit, considerando o idioma inglês [Poświata and Perełkiewicz 2022]. O DepRoBERTa foi usado em uma competição que tinha como desafio classificar postagens do Reddit em uma das seguintes classes: ausente, moderada ou grave. O DepRoBERTa foi ajustado para a tarefa em questão utilizando o conjunto de dados de treinamento e avaliação disponibilizado pela competição. Os experimentos mostraram que a avaliação utilizando uma combinação dos modelos RoBERTa large e DepRoBERTa apresentou o melhor resultado em termos de *F1-score*.

Alguns trabalhos desenvolveram soluções envolvendo a criação de *corpus* e modelo de linguagem no idioma português brasileiro, para um domínio geral de dados. O BERTimbau, desenvolvido por [Souza et al. 2020], é um modelo de linguagem baseado no BERT pré-treinado a partir do *corpus* brWaC (*Brazilian Web as Corpus*) [Wagner Filho et al. 2018], uma coleção de textos de páginas web com 145 milhões de sentenças no idioma português brasileiro. O BERTimbau foi avaliado em três tarefas de PLN, a saber: Similaridade Textual de Sentenças, Reconhecimento de Implicação Textual e Reconhecimento de Entidades Nomeadas.

[Santos et al. 2023] construíram um *corpus* denominado SetembroBR, para ser utilizado no desenvolvimento de modelos preditivos para detecção de depressão. O *corpus* abrange postagens em português extraídas do Twitter (hoje denominado X) de usuários que relataram terem sido diagnosticados com depressão e ansiedade, ou seja, com base em autorrelatos. O trabalho priorizou o *corpus* como sua principal contribuição e destacou sua utilidade para o desenvolvimento de soluções que possam sinalizar evidências da referida doença antes de seu agravamento. Alguns modelos treinados para identificação de depressão e ansiedade com base no SetembroBR foram avaliados com fins de ilustrar seu potencial.

O trabalho de [Costa et al. 2023] desenvolveu o BERTabaporu, um modelo de linguagem BERT pré-treinado com um *corpus* de dados do Twitter, contendo textos de tópicos relacionados a política, saúde mental e Covid-19. Especificamente, para a tarefa de predição de estado de saúde mental (depressão/ansiedade), o modelo foi ajustado utilizando o *corpus* SetembroBR com um esquema de rotulação baseado no autorrelato de depressão dos usuários. Os resultados apontaram que o BERTabaporu conseguiu superar o BERTimbau nas tarefas de classificação de textos em termos das medidas precisão, revocação e *F1*.

Comparando os trabalhos mencionados com a presente proposta, esta se apoia na construção de três artefatos baseados na premissa de que modelos pré-treinados de domínio específico podem trazer resultados mais precisos do que aqueles de domínio geral. São eles: (i) A criação do *corpus* DepreRedditBR no idioma português brasileiro; (ii) Um modelo de linguagem pré-treinado para o domínio da depressão no idioma português brasileiro; e (iii) Um classificador baseado no modelo DepreBERTBR para categorizar postagens considerando três níveis possíveis de depressão: ausente, moderada ou grave.

Particularmente, com respeito ao item (i), este trabalho, assim como o SetembroBR, foca em postagens no português brasileiro mas não obrigatoriamente naquelas já com autorrelatos de usuários que indiquem o rótulo da depressão. A ideia do *corpus* aqui apresentado é que ele contenha textos com possibilidades do possível rótulo e, assim, sirva como um rico vocabulário para o pré-treinamento do modelo. No tocante ao item (ii), apesar dos trabalhos de [Souza et al. 2020] e [Costa et al. 2023] terem desenvolvido modelos pré-treinados no idioma português brasileiro, ambos são de domínio geral, enquanto a abordagem DepreBERTBR foca em dados textuais apenas relacionados à depressão.

4. Abordagem DepreBERTBR

O modelo DepreBERTBR proposto está inserido no contexto particular da depressão. Por ser um modelo especializado, ele pode ser ajustado para utilização em tarefas específicas de PLN, neste caso, para o problema de classificação de postagens com teor depressivo em uma das seguintes classes: ausente, moderada ou grave. A Figura 1 ilustra uma visão geral do processo de desenvolvimento do modelo DepreBERTBR. Cada etapa do processo é descrita a seguir.

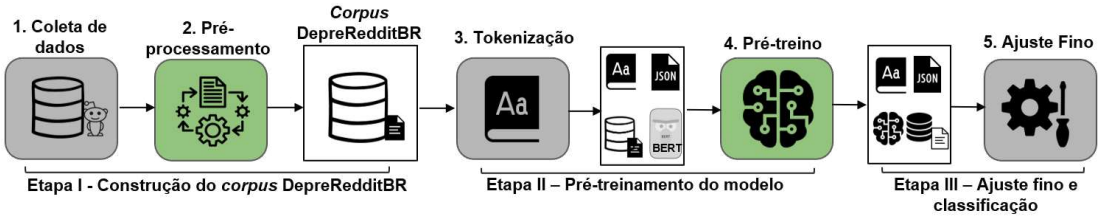


Figura 1. Concepção do DepreBERTBR

4.1. Construção do *corpus* DepreRedditBR

Na Etapa I, objetivou-se a construção do *corpus* nomeado DepreRedditBR. Para isso, foi realizado um levantamento buscando identificar as redes sociais que oferecessem ambientes de discussão relacionadas ao tópico *depressão*. O Reddit demonstrou grande potencial nesse sentido, pois existem subcomunidades (subreddits) ativas que discutem temas afins ao domínio da depressão (e.g., r/arco_iris, r/desabafos, r/EuSouOBabaca, r/BissexualidadeBr, e r/AnsiedadeDepressao). Para realizar a coleta de dados, foi desenvolvida uma aplicação de extração de postagens e/ou comentários de caráter depressivo, conforme subcomunidades alvo e termos especificados a partir dos trabalhos de [Azam et al. 2021] e [da Silva Nascimento et al. 2018]. Os termos foram validados por especialistas do domínio (médicos). Uma amostra deles inclui: *"deprê OR ansiedade OR chorar OR morrer OR matar OR medo OR crises OR desanimado"*. Tendo em vista que os títulos das postagens no Reddit podem conter até 300 caracteres, considerou-se que tais textos também representariam instâncias no conjunto de dados DepreRedditBr. Assim, ao todo, foram coletadas 200.030 instâncias constituídas por títulos, postagens e/ou comentários no idioma português brasileiro.

Haja vista que o pré-treinamento de um LLM requer um *corpus* de entrada com grande volume de dados textuais [Caseli and Nunes 2023], foram examinadas alternativas para aumentar o tamanho do DepreRedditBR. Apesar do *corpus* SetembroBR

[Santos et al. 2023] emergir como uma proposta no contexto da depressão para essa finalidade, não foi possível utilizá-lo devido a restrições determinadas pela política de privacidade da rede social X, a qual não permite compartilhar o conteúdo de suas postagens em repositórios públicos. Adicionalmente, como o SetembroBR apenas disponibiliza os IDs das postagens, tornou-se inviável reproduzir todo o processo de coleta, ainda mais que, a partir do ano de 2023, a plataforma restringiu substancialmente o uso de sua API (*Application Programming Interface*) para coleta de dados em sua rede como uma forma de instigar o uso da versão paga.

Posto isto, foram integrados ao DepreRedditBR dados textuais de um *corpus* disponibilizado na plataforma Kaggle[2], também com postagens do Reddit relacionados à depressão no idioma português brasileiro, constituído por 3.404 postagens. Apesar do acréscimo dos dados citados, em termos quantitativos, o modelo DepreBERTBR necessitava de mais conteúdo textual. Deste modo, vislumbrou-se incorporar conjuntos de dados pertencentes ao domínio da depressão, utilizados em trabalhos relacionados, com postagens nativas no idioma inglês. Para utilizar os conjuntos de dados no idioma inglês seria preciso realizar sua tradução para o português brasileiro, entretanto, realizar a tradução manualmente de uma grande quantidade de textos seria muito custoso e inviável. Sendo assim, foi desenvolvida uma segunda ferramenta voltada à tradução automática de conjuntos de dados do idioma inglês (disponíveis em maior escala) para o português do Brasil. A ferramenta realiza chamadas à API do Google Translate para realizar as traduções.

Particularmente, é importante destacar que embora as soluções de software para tradução automática tenham evoluído ao longo dos anos, ainda assim apresentam limitações que podem influenciar na qualidade da tradução, tais como: (a) perda de significado na tradução de nuances culturais; (b) erros gramaticais e de sintaxe; (c) conotações distintas de palavras nos idiomas envolvidos; e (d) produção de ambiguidade que dificulte a interpretação do texto. Mitigar os problemas apontados exige investimento de tempo, recurso e esforço humano para revisar manualmente uma amostra significativa dos dados traduzidos, ou então experimentar LLMs especializados na tradução automática de dados. Apesar dos ruídos que podem ser provocados pelas limitações existentes nas traduções automáticas, compreende-se que ainda assim esta oferece uma opção plausível, haja vista que o DepreBERTBR contempla uma parcela de dados extraída originalmente de postagens nativas do idioma português brasileiro. Entretanto, medidas podem ser adotadas futuramente para aprimorar a qualidade das traduções automáticas.

O processo de tradução foi realizado em dois *corpora*: o primeiro *corpus* é proveniente do trabalho de [Low et al. 2020], que reuniu postagens de subreddits referentes a transtornos mentais como ansiedade, depressão e suicídio; o segundo *corpus*, disponibilizado no Kaggle[3], continha postagens dos subreddits r/depression e r/SuicideWatch. Ao final, foram integrados ao DepreRedditBR, 338.139 postagens do Reddit no idioma inglês, traduzidas para o idioma português brasileiro. Após realização de rotinas de pré-processamento (e.g., remoção de duplicatas, remoção de emojis e emoticons, remoção de quebras de linhas e marcações especiais), o DepreRedditBR culminou com um total de 509.675 de instâncias, compostas de títulos, postagens e comentários. Cabe ressaltar que a remoção de emojis e emoticons se justifica pelo propósito de produzir um *cor-*

[2]https://www.kaggle.com/datasets/luizfmatos/reddit-portuguese-depression-related-submissions
[3]https://www.kaggle.com/datasets/xavrig/reddit-dataset-rdepression-and-rsuicidewatch

pus unicamente textual. Manter os emojis e emoticons implicaria, a princípio, na sua substituição por expressão/palavra equivalente (e.g., triste, chorando, alegre), o que exigiria um estudo mais dedicado sobre quais termos seriam escolhidos e como poderiam resguardar o sentido da mensagem original. Como esse estudo não faz parte do escopo do trabalho, optou-se por abordá-lo em um momento futuro. Por fim, o *corpus* DepreRedditBR resultante pode ser acessado em https://zenodo.org/records/12761179. A Tabela 1 ilustra uma pequena amostra do *corpus* DepreRedditBR, com textos selecionados aleatoriamente. Percebe-se a evidência de relatos depressivos que transmitem sentimentos de baixa estima, negatividade, desmotivação e uso de medicamentos.

Tabela 1. Amostra do *corpus* DepreRedditBR.

Texto
Desejar a validação dos outros e rejeitar imediatamente qualquer coisa positiva que as outras pessoas digam sobre mim e um tipo especial de inferno. Não tenho confiança em mim mesmo, especialmente em relação a minha aparência física, por isso muitas vezes procuro nos outros coisas sobre as quais posso ser positivo.
Ligando para a linha de socorro enquanto morava em casa? Como? Como posso? Eu realmente não posso pagar uma terapia ou algo assim, então este e meu ultimo recurso. Mas moro em casa e não posso sair em publico para isso
Parei de tomar meus remédios. O que devo fazer? Há cerca de um mês, parei de tomar meus remédios (estou tomando remédios para ansiedade, depressão, enxaquecas, sob e algumas outras coisas). fiquei sem motivação e continuei esquecendo e não tinha vontade, então comecei a toma-los com menos frequência e não tomei nenhum nas ultimas duas semanas.

4.2. Pré-treinamento do Modelo DepreBERTBR

Para o pré-treinamento de um LLM é necessário que o *corpus*, depois de pré-processado, passe pelo procedimento de tokenização (Figura 1, Etapa II). Na atividade de tokenização, o texto é dividido em unidades chamadas de *tokens*. Cada token pode ser uma palavra ou subpalavra, o qual é associado a um identificador (ID) único no vocabulário criado. O DepreRedditBR, *corpus* resultante do final da Etapa I, foi utilizado para pré-treinar um tokenizador com textos do domínio da depressão no idioma português brasileiro. Tendo em vista que estamos diante de um *corpus* especializado no domínio da depressão, no idioma português brasileiro, torna-se necessário gerar o próprio tokenizador para aprender a relação entre as palavras do domínio em particular. Algumas definições usadas na configuração do tokenizador incluem: (a) vocabulário com suporte de até 99.999 palavras; (b) sentença de no máximo 512 tokens; e (c) preparação da lista de tokens de controle (e.g., [MASK], [UNK], [CLS]). O token [MASK] é utilizado para mascarar alguns tokens em uma sentença, enquanto o token [CLS] indica início da sentença de entrada durante o pré-treino do BERT. A configuração do tamanho do vocabulário foi definido para que ele fosse rico na diversidade de palavras, buscando evitar que muitos tokens desconhecidos ([UNK]) fossem gerados durante o processo de tokenização. O tamanho máximo da sentença seguiu a configuração padrão do BERT, que recebe uma sentença com no máximo 512 tokens [Devlin et al. 2019]. Ao final da atividade de tokenização, obtem-se como artefatos resultantes o *corpus* DepreRedditBR tokenizado, um vocabulário gerado a partir desse *corpus* e um arquivo de configuração com a lista de tokens de controle essenciais para o treinamento do BERT. Esses elementos são entradas para a atividade de pré-treino do modelo.

Ainda na Etapa II (Figura 1), foram consideradas as seguintes definições para pré-treinamento do DepreBERTBR: (a) instanciação na versão BERT Base, versão leve do BERT (menor exigência de recursos computacionais) e configurado apenas para a tarefa *Masked Language Modeling*; (b) divisão do *corpus* DepreRedditBR em conjuntos

de treinamento (80%) e teste (20%), este último usado para avaliação do modelo na tarefa *Masked Language Modeling*; (c) nº de épocas de treinamento = 10, ou seja, número de vezes em que o modelo analisa todo o conjunto de dados; (d) taxa de aprendizado padrão do BERT de 5e-5; e (d) estratégia de avaliação em *steps*, ocorrendo a cada 5.000 passos. O pré-treinamento do DepreBERTBR foi realizado em 4 (quatro) dias, utilizando como ambiente computacional a plataforma Google Colaboratory Pro+ e uma GPU NVIDIA Ampere A100 Tensor Core.

4.3. Ajuste Fino e Classificação

Após o pré-treinamento, o DepreBERTBR foi ajustado para uma tarefa de classificação de texto (Etapa III da Figura 1). Esta técnica, denominada de ajuste fino, reutiliza o conhecimento de um modelo pré-treinado para realizar uma tarefa de destino. O conjunto de dados utilizado para ajuste fino do DepreBERTBR foi criado por [Sampath and Durairaj 2022], o qual consiste de postagens do Reddit rotuladas por especialistas de domínio, no idioma inglês. O conjunto de dados contém 10.230 postagens, distribuídas nas seguintes classes: (a) ausente: 34,16%; (b) depressão moderada: 56,39%; e (c) depressão grave: 9,45%.

O conjunto de dados usado para treino e teste da tarefa preditiva também foi traduzido utilizando a ferramenta de tradução usada na Etapa I da Figura 1, visando obter uma versão correspondente dos dados no idioma português brasileiro. Vale ressaltar que esse conjunto de dados também foi subordinado às rotinas de pré-processamento que ocorreram na construção do DepreRedditBR (Seção 4.1). A Tabela 2 mostra um fragmento do *corpus* pré-processado usado para treino e teste do classificador. A coluna $classe$ representa o rótulo da postagem referente ao conteúdo da coluna $texto$. Para avaliar o desempenho do classificador foram realizadas 10 iterações executando o ajuste fino do DepreBERTBR. Em cada iteração, foi utilizado um subconjunto do *corpus* da tarefa de classificação, sendo cada subconjunto dividido em conjunto de dados de treino e teste e tokenizado utilizando o tokenizador do DepreBERTBR (Seção 4.2).

Tabela 2. Amostra do *corpus* rotulado.

Texto	Classe
Alguém quer só conversar?	ausente
Estou em uma festa de ano novo e quero chorar, estou tendo um ataque de pânico em motivo algum, não quero mais estar aqui	moderada
Eu não quero morrer, só quero parar de viver. Isso faz sentido? Deus, eu odeio minha vida	grave

5. Experimentos e discussão

Os experimentos foram planejados de maneira a realizar um ajuste fino para a tarefa de classificação de texto utilizando o *corpus* rotulado por [Sampath and Durairaj 2022], conforme descrito na Seção 4.3. O problema consiste em classificar postagens de redes sociais do conjunto de teste quanto ao nível de severidade de depressão em uma das três classes: ausente, moderada ou grave. O experimento compara os modelos DepreBERTBR, BERTimbau e BERTabaporu. Os modelos selecionados como *baselines* foram escolhidos por serem modelos pré-treinados com *corpora* no idioma português brasileiro. Ainda, o BERTabaporu incluiu em seu treinamento dados relacionados à depressão e transtorno de ansiedade. Para o ajuste fino, tanto o DepreBERTBR, quanto os modelos utilizados como

baselines, foram configurados para utilizar um otimizador Adamw, com taxa de aprendizado igual a 5e-5, valor padrão para o BERT instanciado utilizando a biblioteca *hugging face*[4] e função de ativação softmax na camada de saída. A Tabela 3 sumariza informações referentes ao *corpus* usado no pré-treinamento de cada modelo.

Tabela 3. Características do *corpus* utilizado para pré-treinamento dos modelos.

Modelo	Domínio	Origem	Tamanho do corpus	(%) de dados saúde/saúde mental	Vocabulário
DepreBERTBR	Depressão	Reddit	509.189 mil	100%	99.999 mil
BERTimbau	Geral	Web	145 milhões	6%	30.000 mil
BERTabaporu	Geral + Saúde mental	Twitter	238 milhões	3,8 %	64.000 mil

A Tabela 4 apresenta o resultado da classificação geral para os três modelos comparados, tanto por iteração quanto na média obtida em termos de *F1*. As medições foram realizadas considerando o treinamento dos modelos nos cenários com 2 e 10 épocas. Os experimentos por quantidade de épocas no treinamento têm o intuito de examinar o desempenho dos modelos quando incrementado o número de ciclos de treinamento. Nota-se que a diferença média de *F1* dos modelos comparados é muito sutil em ambos os cenários avaliados, sugerindo que não há diferença significativa na média de *F1* dos modelos. As medições de *F1* por iteração revelam que os modelos vão melhorando o aprendizado do problema gradativamente até a 5ª iteração, maximizando, de forma consolidada, o desempenho das classificações da 6ª iteração em diante.

Tabela 4. F1-score por iteração no treinamento com 2 e 10 épocas.

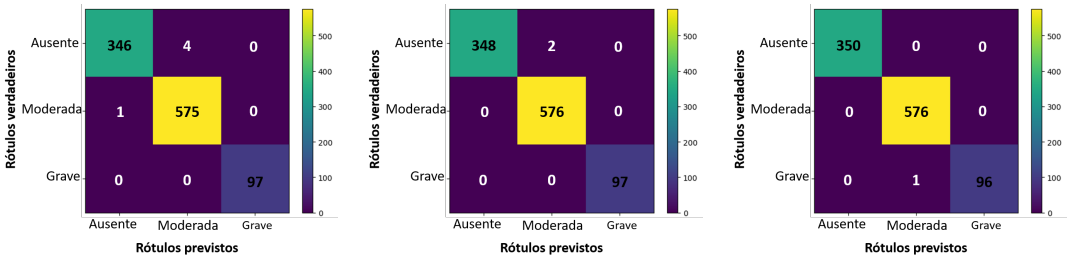
Iteração	**2 épocas**			**10 épocas**		
	DepreBERTBR	BERTimbau	BERTabaporu	DepreBERTBR	BERTimbau	BERTabaporu
1	0,50	0,55	0,54	0,49	0,59	0,51
2	0,68	0,67	0,67	0,69	0,76	0,64
3	0,80	0,74	0,81	0,75	0,87	0,79
4	0,93	0,85	0,94	0,88	0,92	0,94
5	0,97	0,92	0,98	0,95	0,94	0,96
...	...	...	...	...	...	...
10	**0,99**	**0,99**	**0,99**	**0,99**	**0,99**	**0,99**
avg(F1)	0,89	0,86	0,89	0,87	0,90	0,88

Quando o treinamento é realizado no cenário com 10 épocas, observa-se uma tendência de melhoria de *F1* do BERTimbau e BERTabaporu nas iterações iniciais, equilibrando o desempenho entre os modelos comparados na segunda metade das iterações. Supõe-se, então, que os modelos BERTimbau e BERTabaporu, por serem pré-treinados com uma massa de dados muito superior ao usado pelo DepreBERTBR, tendem a apresentar melhor precisão nas iterações iniciais quando submetidos a um número maior de ciclos de treinamento, mesmo sendo modelos pré-treinados com dados de domínio geral.

A Figura 2 fornece uma visão geral do desempenho dos modelos quanto aos erros e acertos das classificações, considerando o treinamento com 10 épocas. A matriz de confusão foi coletada na 10ª iteração. Nessa iteração, percebe-se que todos os modelos apresentaram um notável desempenho na predição das três classes do problema. O BERTimbau e o DepreBERTBR conseguiram classificar corretamente quase a totalidade dos

[4]https://huggingface.co/

exemplos das classes *depressão moderada* e *depressão grave*. O BERTabaporu demonstrou um destacado desempenho na predição das três classes. Apenas alguns exemplos da classe *ausente* foram classificados incorretamente como *depressão moderada* pelos modelos DepreBERTBR e BERTimbau. Com um maior número de épocas no treinamento, os modelos tendem a refinar ainda mais seus acertos quanto à classificação do grau da depressão. Importante destacar o excelente desempenho dos modelos ao classificar corretamente instâncias da classe minoritária (*depressão grave*), o que na prática é um problema normalmente a ser tratado quando há um desbalanceamento de classes em um conjunto de dados. Ressalta-se também que as últimas iterações, analisadas de forma isolada, apresentam um desempenho próximo de 100% de acerto, porém, a média de *F1*, ao longo das 10 iterações, está entre [0,86...0,90], justamente porque os modelos apresentam um desempenho inferior nas iterações iniciais.



(a) DepreBERTBR (b) BERTimbau (c) BERTabaporu.

Figura 2. Matriz de confusão para a última iteração.

De acordo com os resultados obtidos, as QPs definidas neste trabalho são respondidas da seguinte maneira:

(i) Com respeito à **QP1**, a abordagem DepreBERTBR inclui uma solução para classificação de postagens de redes sociais para três níveis de depressão. A solução apresentada e avaliada é baseada no uso de LLMs, e o resultado obtido indica um desempenho promissor para textos no idioma português brasileiro. Os resultados de *F1-score* e a matriz de confusão da Figura 2 fornecem subsídios para essa resposta.

(ii) No tocante à **QP2**, os resultados dos experimentos trazem evidências de que o pré-treinamento de um LLM com um *corpus* específico do domínio da depressão é favorável para alcançar resultados promissores em problemas de classificação do nível de severidade de depressão. A Tabela 3 constata que o BERTimbau e BERTabaporu realizaram um pré-treino com milhões de textos/sentenças, o que impacta substancialmente no tempo necessário para pré-treinamento, infraestrutura computacional e tamanho do modelo. O processamento dessas massas de dados requer recursos computacionais que, geralmente, só estão disponíveis em serviços de plataformas de dados na nuvem, sendo necessários vários dias ou até semanas para realizar testes e experimentação. Os custos elevados para manter esses recursos podem ser empecilhos para a realização e implementação de pesquisas e aplicações, como no caso do pré-treinamento de modelos como o BERTimbau e BERTabaporu. Apesar do tamanho do *corpus* DepreRedditBR ser consideravelmente inferior, o DepreBERTBR mostrou-se competitivo em comparação com os modelos BERTimbau e BERTabaporu no cenário de classificação proposto. Logo, o DepreBERTBR demonstra potencialidade para ajustar-se a problemas de classificação

de nível de severidade de depressão.

6. Considerações e trabalhos futuros

Este trabalho apresentou uma abordagem para classificação de postagens de redes sociais com tendências depressivas, que se apoiou na construção de um *corpus* (DepreRedditBR) e de um modelo de linguagem pré-treinado (DepreBERTBR), considerando o idioma português brasileiro. O DepreBERTBR foi ajustado para a tarefa de classificação de postagens conforme três níveis de depressão: ausente, moderada ou grave.

Os resultados obtidos na avaliação experimental demonstram que a diferença média de *F1* dos modelos comparados (BERTAbaporu e BERTimbau) é muito sutil tanto no cenário avaliado com 2 épocas quanto no cenário com 10 épocas, sugerindo que não há diferença significativa na média de *F1* dos modelos. As medições de *F1* por iteração revelam que os modelos vão melhorando o aprendizado do problema gradativamente até a 5ª iteração, maximizando, de forma consolidada, o desempenho das classificações da 6ª iteração em diante. Nota-se assim que, quando há mais dados para o treinamento (BERTimbau e BERTApabaporu) e mais tempo para a convergência do aprendizado, tende-se a ter resultados mais assertivos. Por outro lado, é importante salientar que, apesar do tamanho do *corpus* DepreRedditBR ser consideravelmente inferior aos demais comparados, o DepreBERTBR conseguiu alcançar resultados competitivos na tarefa de classificação em relação aos modelos comparados, também pré-treinados no idioma português do Brasil. O DepreBERTBR apresentou uma média de *F1*=0,88, com tendência a maximização de seu desempenho preditivo nas últimas iterações incrementais de treino e teste do classificador.

Alguns trabalhos futuros planejados são: (i) Expansão do corpus DepreRedditBR com mais postagens no idioma português brasileiro no domínio da depressão; (ii) Adoção de estratégias para mitigação dos ruídos produzidos pelas traduções automáticas visando aprimorar a qualidade das traduções, de tal forma que o resultado da tradução automática possa ser combinado com algum procedimento científico de validação humana e/ou aplicação de técnicas de PLN avançadas; (iii) Realização de novo pré-treinamento do modelo DepreBERTBR, utilizando como base a versão maior do BERT (BERT Large) e o *corpus* aumentado; e (iv) Implementação do ajuste fino do DepreBERTBR para a tarefa de classificação de textos (no português brasileiro), de acordo os quatro níveis de depressão presumidos no Inventário de Depressão de Beck, a saber: ausente, leve, moderada ou grave.

Referências

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Azam, F., Agro, M., Sami, M., Abro, M. H., and Dewani, A. (2021). Identifying depression among twitter users using sentiment analysis. In *2021 international conference on artificial intelligence (ICAI)*, pages 44–49. IEEE.

Cacheda, F., Fernandez, D., Novoa, F. J., Carneiro, V., et al. (2019). Early detection of depression: social network analysis and random forest techniques. *Journal of medical Internet research*, 21(6):e12554.

Caseli, H. d. M. and Nunes, M. d. G. V. (2023). *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*. BPLN, 2a 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.

da Silva Nascimento, R., Parreira, P., dos Santos, G. N., and Guedes, G. P. (2018). Identificando sinais de comportamento depressivo em redes sociais. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*. SBC.

de Psiquiatria, A. A. (2022). *Manual Diagnóstico e Estatístico de Transtornos Mentais - DSM-5-TR*. Artmed.

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.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., and Cambria, E. (2022). MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190. European Language Resources Association.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., and Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Oliveira, B. S. N., do Rêgo, L. G. C., Peres, L., da Silva, T. L. C., and de Macêdo, J. A. F. (2022). Processamento de linguagem natural via aprendizagem profunda. *Sociedade Brasileira de Computação*.

OMS (2023). Organização mundial de saúde (oms): Desordem depressiva (depressão). https://www.who.int/news-room/fact-sheets/detail/depression. Último Acesso 28 de Mai 2024.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

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.

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.

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.

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)*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.