DepreReddit BR: Um conjunto de dados textuais com postagens depressivas no idioma português brasileiro

Ayrton Douglas Rodrigues Herculano[1], Taw-Ham Almeida Balbino de Paula[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, balbino.taw}@academico.ifpb.edu.br, {damires,alex}@ifpb.edu.br

Abstract. *Depression is a mental disorder that often presents disabling characteristics. Monitoring users activity on their social networks can help in the early identification of depression. Research has sought textual data to train models and generate computational solutions, but the majority still uses data in the English language. In this light, this work has built DepreRedditBR, a textual data set with 509,675 instances of posts with depressive content from the Reddit network in Brazilian Portuguese. DepreRedditBR was used to pre-train a LLM, whose acquired knowledge allowed the model, after tuning, to classify posts according to the degree of depression.*

Resumo. *A depressão é um transtorno mental que apresenta características, muitas vezes, incapacitantes. O monitoramento da atividade de usuários em suas redes sociais pode ajudar na identificação precoce da depressão. Pesquisas tem buscado dados textuais para treinar modelos e gerar soluções computacionais, porém a maioria ainda utiliza dados no idioma inglês. Neste cenário, este trabalho construiu o DepreRedditBR, um conjunto de dados textuais com 509.675 instâncias de postagens com teor depressivo a partir da rede Reddit no idioma português brasileiro. O DepreRedditBR foi utilizado para o pré-treinamento de um LLM, cujo conhecimento adquirido permitiu que o modelo, depois de ajustado, classificasse postagens de acordo com o grau de depressão.*

1. Introdução

A depressão é um transtorno mental que afeta a natureza emocional, psicológica e física de um indivíduo. Alguns sinais e sintomas que sinalizam um padrão de suspeita de desenvolvimento da depressão incluem [WHO 2023]: humor deprimido, perda de prazer ou interesse em realizar atividades, sentimento de culpa, baixa autoestima e insônia. Há também fatores de risco que podem contribuir com o desenvolvimento da depressão, tais como histórico familiar, excesso de peso, luto, separação conjugal, entre outros.

De acordo com a Organização Mundial de Saúde (OMS), no ano de 2022, cerca de 280 milhões de pessoas foram acometidas pelo transtorno da depressão em todo o mundo. A previsão da OMS é que, até 2030, ela será a doença mais comum mundialmente [WHO 2023]. Entretanto, detectar depressão é uma tarefa complexa, visto que sua identificação decorre da combinação de sinais e sintomas que podem persistir diariamente por no mínimo duas semanas [Nardi et al. 2021]. Conforme [Cacheda et al. 2019],

a prevenção contra a depressão pode acontecer de diferentes formas, sendo uma delas por meio do monitoramento do comportamento de usuários em redes sociais.

As redes sociais tornaram-se um espaço virtual muito popular entre a maioria das pessoas, podendo ser utilizadas para várias finalidades como, por exemplo, divulgar serviços e produtos, expressar sentimentos, compartilhar experiências e relatar atividades do cotidiano. As interações dos usuários em postagens nas redes sociais evidenciam uma linguagem que pode revelar emoções e sentimentos dos mais variados, tais como inutilidade, solidão, abandono e ódio a si próprio, características estas típicas da depressão [De Choudhury et al. 2013]. Em alguns casos, usuários se sentem mais à vontade para demonstrar sentimentos reprimidos quando interagem em grupos com temáticas comuns à depressão e ansiedade. Isso tem acontecido especialmente em redes sociais baseadas em comunidades como a Reddit [Uban et al. 2021]. A rede social Reddit[1] possui comunidades, chamadas de *subreddits*, que são espaços onde os usuários com interesses comuns postam sobre assuntos específicos como, por exemplo, ansiedade [Ji et al. 2022], seja por meio de conteúdo em formato de texto, imagem ou links [Pérez et al. 2022].

Pesquisas nas áreas de Psiquiatria, Psicologia, Sociolinguística e Neurociência, associadas a técnicas computacionais específicas como a Análise de Sentimentos e o Processamento de Linguagem Natural (PLN), buscam aprimorar a compreensão da relação entre o comportamento das pessoas, seus sentimentos e emoções, utilizando como fontes de dados textos postados em redes sociais [Vedula and Parthasarathy 2017, Cacheda et al. 2019]. No entanto, uma revisão sistemática da literatura (RSL), realizada por [Herculano et al. 2022], evidenciou que a grande maioria desses trabalhos utiliza conjuntos de dados no idioma inglês, existindo poucos trabalhos com dados sobre transtornos mentais no idioma português brasileiro. Alguns trabalhos de pesquisa começaram a investir nestas demandas por dados textuais em português [Santos et al. 2023, Sperling and Ladeira 2019, da Silva Nascimento et al. 2018], mas ainda há poucos conjuntos de dados com grande quantidade de postagens oriundas de redes sociais relacionadas ao domínio da depressão no idioma português brasileiro.

Diante do cenário exposto, este trabalho apresenta a sistemática usada para criação de um conjunto de dados textuais, com teor depressivo, a partir de postagens do Reddit no idioma português brasileiro. Denominado de `DepreRedditBR`, o conjunto de dados foi utilizado para o pré-treinamento de um modelo de linguagem em larga escala (do inglês, *Large Language Model* ou LLM) especializado ao domínio da depressão. LLMs são modelos de linguagem treinados com uma enorme quantidade de dados [Caseli and Nunes 2023]. Posteriormente, com o conhecimento aprendido a partir do conjunto de dados `DepreRedditBR`, o modelo pré-treinado foi ajustado à tarefa de classificação de texto, e avaliado quanto ao seu desempenho em identificar corretamente postagens em um dos seguintes graus de depressão: ausente, moderada ou grave.

O restante deste trabalho está organizado da seguinte maneira: A Seção 2 apresenta alguns trabalhos relacionados. A Seção 3 descreve as etapas seguidas para a construção do `DepreRedditBR`. A Seção 4 aborda algumas utilizações do conjunto de dados construído. Por fim, a Seção 5 tece considerações finais e indica trabalhos futuros.

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

2. Trabalhos Relacionados

Alguns trabalhos encontrados na literatura propõem a construção de conjuntos de dados textuais contendo postagens associadas a sinais de possível depressão [da Silva Nascimento et al. 2018, Low et al. 2020, Sampath and Durairaj 2022, Santos et al. 2023], conforme descritos a seguir.

O `Depress-pt-br` é um conjunto de dados idealizado por [da Silva Nascimento et al. 2018] constituído por termos referentes à depressão no idioma português brasileiro. Seu processo de construção contemplou etapas de coleta de dados, pré-processamento e integração com dicionários léxicos. Na primeira etapa, foram coletados textos com teor depressivo a partir de postagens provenientes do website de perguntas e respostas *Yahoo Respostas*, descontinuado no ano de 2021, utilizando *strings* de busca tais como "depressão", "depressão+suicídio" e "depressão+socorro". Para os 3.000 documentos resultantes, foram extraídos o título, o detalhe (informações adicionais da pergunta) e a melhor resposta de cada pergunta. Na segunda etapa, realizou-se um pré-processamento dos dados para tokenização do texto, normalização dos tokens para caracteres minúsculos e remoção de pontuações. Na última etapa, os dados pré-processados foram unificados com os dicionários léxicos ANEW-br [Kristensen et al. 2011] e *Linguistic Inquiry and Word Count* LIWC2007-pt-br [Balage Filho et al. 2013] para reconhecer entradas positivas e negativas. Além da valência positiva e negativa de cada termo, os pesquisadores adicionaram uma categoria chamada *depress*, para indicar palavras relacionadas à depressão.

O conjunto de dados proposto por [Low et al. 2020], denominado `Reddit Mental Health Dataset`, contém postagens relacionadas ao âmbito de transtornos mentais, no idioma inglês, a partir da rede social Reddit. Um dos objetivos para concepção do referido conjunto de dados foi o de prover subsídios para avaliar como a covid-19 impactou a saúde mental das pessoas acometidas com esses transtornos. Para isso, os autores coletaram postagens de 15 *subreddits* associados a temas sobre saúde mental como, por exemplo, *r/depression*, *r/anxiety*, *r/SuicideWatch* e *r/schizophrenia*. Também foram coletadas postagens de 11 subreddits que não estavam relacionados à saúde mental como, por exemplo, *r/fitness*, *r/meditation*, *r/parenting*, *r/personalfinance* e *r/relationships*, utilizados para análises baseadas em "modelagem de tópicos". O `Reddit Mental Health Dataset` agrega postagens compreendidas em dois períodos: antes e durante a pandemia de Covid-19. Para avaliar o impacto provocado pela pandemia em diferentes comunidades (*subreddits*) associadas à saúde mental, os autores aplicaram técnicas de PLN conjuntamente com métodos de aprendizado de máquina.

[Sampath and Durairaj 2022] construíram um conjunto de dados para ser utilizado em problemas de classificação de depressão a partir de postagens publicadas no Reddit no idioma inglês. As postagens foram coletadas tendo como alvo os seguintes subreddits: *r/Mental Health*, *r/depression*, *r/loneliness*, *r/stress* e *r/anxiety*. Após a coleta, as postagens passaram por etapas de pré-processamento para remoção de caracteres não enquadrados no padrão ASCII (*American Standard Code for Information Interchange*), links HTML (*HyperText Markup Language*) e emoticons. Além disso, o título e o corpo da postagem foram combinados em uma única coluna. O conjunto de dados, então, foi rotulado por dois especialistas de domínio considerando três classes: "ausente", "moderada"ou "grave". Os autores utilizaram o conjunto de dados criado para realizar experimentos de

classificação quanto ao nível de depressão. Para isso, utilizaram algoritmos de aprendizado de máquina como *Random Forest*, *Support Vector Machine (SVM)* e *K-Nearest Neighbour (KNN)*.

O trabalho de [Santos et al. 2023] culminou com a criação do `SetembroBr`, que inclui dois conjuntos de dados em português voltados, respectivamente, para o domínio da depressão e do transtorno de ansiedade. A motivação por trás de sua criação foi a de proporcionar um conjunto de dados que servisse de suporte ao desenvolvimento de modelos preditivos, focados na detecção de depressão e do transtorno de ansiedade. Os dados do `SetembroBr` foram coletados a partir de postagens da rede social Twitter (atualmente chamado de X[2]) no período compreendido entre setembro de 2019 a fevereiro de 2021, considerando os perfis qualificados como "usuários diagnosticados com depressão". Para identificar tais usuários, os autores adotaram a estratégia da rotulação baseada no autorelato nas postagens onde usuários sinalizavam seu diagnóstico de depressão usando expressões como, por exemplo, "Comecei a tomar medicamentos antidepressivos". Alguns perfis de usuários foram descartados na fase de coleta, a saber: (a) perfis que relataram o desenvolvimento de outros transtornos mentais (e.g., borderline, transtorno bipolar, esquizofrenia) e; (b) perfis identificados como "usuários de controle", caracterizados por não apresentaram indícios de sintomas depressivos e/ou de transtorno de ansiedade, ou apresentarem uma atividade de mais de 1.000 tweets postados e possuírem mais de 10.000 seguidores no Twitter. Para avaliar a relevância do conjunto de dados `SetembroBr`, foram desenvolvidos experimentos para treinar modelos capazes de predizer se um conjunto de postagens do usuário apresentava ou não indícios de depressão ou transtorno de ansiedade.

Considerando a perspectiva de conjuntos de dados especializados em postagens de cunho depressivo, no idioma português brasileiro, as opções são ainda limitadas. O `SetembroBr` apresenta uma proposta similar, porém, neste trabalho, são disponibilizados apenas os IDs da postagens porque a política de privacidade do X[3] restringe o compartilhamento do conteúdo das postagens em repositórios públicos. A barreira imposta no contexto do trabalho do `SetembroBr` corroborou com a decisão de produzir um novo conjunto de dados.

O `DepreRedditBR`, em comparação aos trabalhos descritos, tem como diferencial a concepção de um conjunto de dados textuais especializado em postagens específicas do domínio da depressão, no idioma português brasileiro, e disponível para uso. Salienta-se que, os textos coletados não são apenas de pessoas que se declararam diagnosticadas com depressão, ou seja, já estavam acometidos pela doença, mas de textos cujo teor poderia indicar um possível quadro de depressão a partir de termos avaliados e validados por especialistas médicos, após um levantamento da literatura [Herculano et al. 2022]. O conjunto de dados também não possui variável com rótulo, tendo em vista seu uso inicial prioritariamente para o pré-treinamento de um LLM. Assim, o objetivo deste trabalho de construção do conjunto de dados foi coletar textos com possível teor depressivo para que pudessem ser utilizados, inicialmente, no pré-treinamento de LLMs que, após ajustados, pudessem realizar a tarefa de classificação de postagens ou mesmo de perfis de usuários quanto a um possível grau de depressão.

[2]https://x.com/

[3]Políticas definidas a partir do ano de 2023.

3. Metodologia

A concepção do conjunto de dados `DepreRedditBR` foi uma necessidade primária para o pré-treinamento de um LLM especializado no domínio da depressão, com o objetivo de ser ajustado para uma tarefa de classificação quanto ao nível de depressão em uma determinada postagem, com base nas possíveis classes: ausente, moderado e grave. O pré-treino de um LLM normalmente exige um grande volume de dados, facilmente na ordem de milhões de sentenças ou documentos textuais. Um conjunto de dados textuais para essa finalidade tende a produzir um melhor conhecimento do domínio, aliado a outros fatores que também influenciam no desempenho do modelo. Com maior quantidade de dados, um LLM é capaz de capturar o vocabulário com sutilezas da linguagem, entender o contexto e aprender com um vasto número de exemplos. Por outro lado, processar essa massa de dados exige uma infraestrutura de recursos computacionais (memória RAM, cota de disco e poder de processamento) que normalmente não está ao alcance da maioria das instituições acadêmicas do país. A solução, então, é contratar o serviço de uma plataforma de processamento de dados na nuvem, a qual dispõe de máquinas sofisticadas e de grande poder computacional apropriadas para tarefas de aprendizado de máquina. Para se ter uma ideia, o pré-treino de um LLM pode levar muitos dias, dependendo da configuração do serviço de processamento contratado [Souza et al. 2020]. No entanto, o custo para viabilizar a realização de testes e experimentos pode ser motivo de obstáculo para pesquisas e aplicações que possuem restrições orçamentárias e necessitam de resultados rápidos. Motivado principalmente por este atual cenário de pesquisas acadêmicas que utilizam LLMs, produzir um conjunto de dados menor que consiga compreender o domínio do conhecimento para pré-treinamento de um LLM é uma alternativa que pode trazer benefícios em relação ao tempo de processamento e custo computacional.

O conjunto de dados ou *corpus* proposto neste artigo teve como objetivo principal gerar um vocabulário amplo que pudesse servir de base para o pré-treinamento de um LLM focado no domínio da depressão no idioma português brasileiro. A Figura 1 mostra o pipeline para a construção do `DepreRedditBR`. A etapa de coleta reuniu dados da rede social Reddit a partir de três fontes: (1) Postagens e comentários em português extraídos por meio da ferramenta chamada "Extrator"; (2) Conjunto de dados com postagens no idioma português brasileiro obtidos do repositório Kaggle[4], e (3) Textos de postagens no idioma inglês traduzidos para o português brasileiro. Após a coleta, os dados textuais passaram pela etapa de pré-processamento. As duas etapas são descritas a seguir.

3.1. Coleta de dados

A primeira decisão que norteia o início do processo de coleta de dados é identificar qual(is) fontes de informação serão utilizadas para produzir o conjunto de dados. Apoiado na decisão de coletar dados de redes sociais, identificou-se o potencial de explorar a rede social Reddit como provedora de conteúdo. A opção pelo Reddit se justifica pelo fato de haver espaços de discussões denominados de *subreddits*, concentrados em diferentes tópicos, inclusive sobre temáticas relacionadas à depressão. Ainda, enfatiza-se o requisito não-funcional de a rede social dispor de condições de extração de dados por meio de uma API (*Application Programming Interface*) própria, o que permite a extração dos dados de forma viável.

[4]https://www.kaggle.com/datasets/luizfmatos/reddit-portuguese-depression-related-submissions

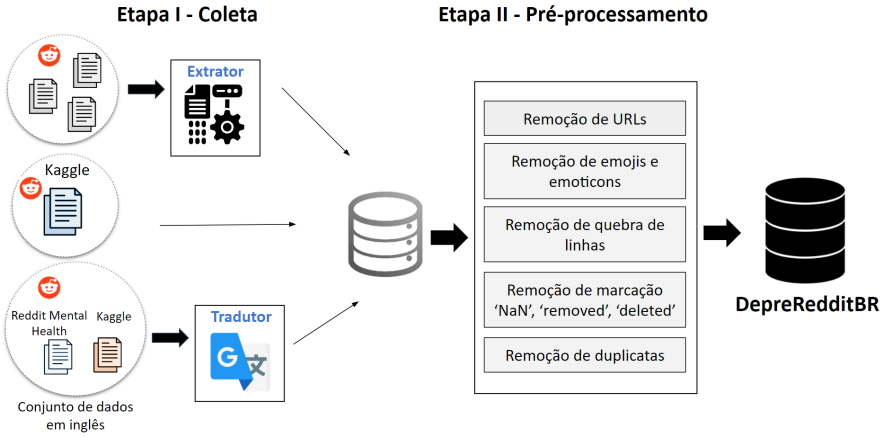

Figura 1. Pipeline de Construção do DepreRedditBR

Para materializar a coleta de dados, foi necessário desenvolver uma ferramenta especializada na extração de dados do Reddit, denominada "Extrator". A ferramenta faz uso da biblioteca PRAW[5] (*The Python Reddit API Wrapper*), a qual simplifica o acesso à API do Reddit. Por meio de uma interface web com o usuário, são informados os parâmetros necessários para iniciar a coleta das postagens via API, a saber: credenciais de acesso, *subreddits* alvo da extração e a string de busca com as palavras-chave que serão usadas para selecionar as postagens desejadas. Uma postagem do Reddit normalmente contém diversos metadados associadas ao seu conteúdo. Para atender aos objetivos deste trabalho, foram considerados os seguintes metadados: (a) título da postagem; (b) corpo do texto da postagem, e (c) comentários da postagem. As coletas de dados foram realizadas entre fevereiro e outubro de 2023, considerando publicações realizadas entre janeiro de 2018 a outrubro de 2023. O `DepreRedditBR` foi publicado em 15 de julho de 2024 e atualizado em 18 de julho de 2024.

As subcomunidades alvo da coleta foram investigadas conforme associação com a temática de saúde mental (particularmente em que a depressão é o foco principal das discussões), identidade e experiência (aquelas que abordam discussões sobre lutas enfrentadas por grupos específicos) e subcomunidades com o objetivo de apoiar pessoas (por exemplo, problemas afetivos ou relacionados como intimidação ou preconceito). Assim, após averiguação pelo grupo de pesquisa das subcomunidades existentes no Reddit, com potenciais postagens dentro do contexto do projeto, as seguintes foram escolhidas: *r/arco_iris*, *r/desabafos*, *r/desabafo*, *r/relacionamentos*, *r/transbr*, *r/EuSouOBabaca*, *r/BissexualidadeBr*, *r/AnsiedadeDepressao*, *r/brasil*, *r/relatosdoreddit* e *r/PsicologiaBR*. As subcomunidades elencadas foram selecionadas de forma empírica, conforme potencial observado nos relatos e comentários de natureza depressiva compartilhados pelos seus usuários. Amostras de postagens e de relatos foram apresentadas aos médicos especialistas do projeto (psiquiatras) que concordaram com a adequação dos conteúdos destas amostras.

Os termos de busca foram selecionados a partir de um levantamento da literatura,

[5]https://praw.readthedocs.io/en/stable/index.html

em especial dos trabalhos de [Azam et al. 2021] e [da Silva Nascimento et al. 2018]. Em seguida, o conjunto de termos foi examinado e boa parte deles foi aprovada pelos especialistas médicos. Os termos aprovados pelos especialistas foram selecionados para a busca, a saber: *"deprê OR ansiedade OR chorar OR morrer OR matar OR medo OR crises OR chorando OR Só OR sozinho OR solidão OR desolado OR desolada OR morto OR vazio OR suicídio OR surto OR surtei OR surtar OR depressivo OR depressiva OR depressão OR depressao OR ansioso OR ansiosa OR desespero OR desesperado OR desesperada OR solitário OR solitária OR solitario OR solitaria OR melancólico OR melancólica OR desânimo OR tristeza OR depresso OR infeliz OR angustiado OR choro OR cortar OR corte OR culpa OR culpado OR culpado OR culpando OR deprimido OR deprimida OR desamparado OR desamparada OR desanimado OR desanimada OR desmotivado OR desmotivada OR doloroso OR dolorosa OR dor OR dores OR frustrado OR insonia OR insônia OR machucado OR morreu OR morte OR noite OR pranto OR prantos OR pulsos OR punicao OR punição OR sangrar OR sangrento OR solidao OR solitario solidão OR solitário OR sozinho OR suicidar OR suicidas OR suicidio OR suicídio OR tedio OR tédio OR triste OR desesperança OR melancolico OR melancolica OR melancolia OR cansado OR cansada OR sufocado OR sufocada"*.

É importante salientar que os termos de busca são palavras que os especialistas destacam como alerta a um possível sinal de sentimento depressivo no texto publicado por um usuário, atuando neste caso como um filtro de seleção de postagens. Isso não garante que o comentário ou postagem foi publicado por um usuário que está sofrendo de depressão, ou que todas as mensagens sejam de fato depressivas, pois isso depende da análise de contexto para dizer que, por exemplo, a ocorrência da palavra "angustiado" em um dado texto faz parte de um relato depressivo ou se refere a outra situação, análise esta que não faz parte do escopo do presente trabalho. Entretanto, a inspeção realizada nos *subreddits* selecionados apresenta, em sua maioria, mensagens de cunho depressivo.

Como resultado da seleção e extração das postagens e comentários, a ferramenta "Extrator" gerou um conjunto de dados em formato CSV (*Comma Separated Values*). A Figura 2 mostra um fragmento do conjunto de dados resultante da coleta, constituído pelas colunas *title* e *text*. Uma vez que o título de uma postagem pode ter até 300 caracteres, este também pode ser utilizado como texto a compor o conjunto de dados. Assim, os títulos, textos das postagens e comentários tornaram-se, cada um, instâncias do conjunto de dados resultante. Ao término dessa coleta por meio da ferramenta "Extrator", o `DepreRedditBR` resultou em um total de 200.030 instâncias.

Considerando que o volume de dados por ora coletado ainda era insuficiente para pré-treinar um LLM conforme relatos da literatura [Caseli and Nunes 2023], foram buscadas alternativas para agregar informações ao conjunto. Uma delas foi aproveitar um conjunto de dados dentro da temática depressão disponibilizado pela plataforma Kaggle, no idioma português brasileiro. Este conjunto de dados possui 3.404 postagens extraídas do Reddit. As postagens disponíveis neste conjunto de dados também se tornaram instâncias de dados na composição do `DepreRedditBR` e incluem o título, o corpo do texto, data e hora da postagem e o *subreddit* em que a postagem foi publicada.

A inclusão dos dados do conjunto do Kaggle no `DepreRedditBR` gerou um impacto muito sutil em termos quantitativos. Devido à escassez de outros conjuntos de dados no idioma português, vislumbrou-se a oportunidade de aproveitar alguns conjuntos de da-

	text	title
0	Parece que ninguem quer ficar perto de mim, e,...	Nao consigo ter boas relações, e isso tem me c...
1	Eu sou deprimido, porem tem períodos que fico ...	Estou vazio e não sei o que fazer
2	Boa tarde. Bem, é basicamente um desabafo. A v...	Depressão e Ansiedade
3	Me tirem uma duvida, ser perfeccionista ao ac...	Passando por isso
4	\nOi gente, desculpa o textão, já mandei isso ...	Não consigo mais viver com ansiedade
...	...	...
90	Olá, ansioso!\n\nQueria saber como você tem li...	Ansiedade
91	Tenho muitas dores de cabeça que são recorrent...	Tenho muitas dores de cabeça
92	Podem ouvir aqui o meu podcast sobre problemas...	Podcast para vocês!
93	NaN	Gostaria de desejar um bom dia à todos os memb...
94	É um medo fora do comum. Já aconteceu crises d...	Eu sinto nada e tudo ao mesmo tempo.

Figura 2. Fragmento do conjunto de dados coletado com a ferramenta Extrator

dos do domínio da depressão disponibilizados na língua inglesa. Para isso, consideraram-se algumas evidências relatadas pela RSL realizada por [Herculano et al. 2022], que apresentava trabalhos referentes à detecção de indícios de depressão utilizando dados no idioma inglês. Porém, para tornar isso possível, seria necessário desenvolver uma solução mínima de tradução em massa de textos do idioma inglês para o português-brasileiro.

Para tal solução, nitidamente não seria viável em termos de custo e tempo realizar as traduções com a precisão do esforço humano. Dessa forma, foi utilizada a API do Google Translate para realizar a tarefa de tradução do inglês para o português brasileiro de dois conjuntos de dados coletados do Reddit. O primeiro, proveniente do trabalho de [Low et al. 2020], reúne postagens de *subreddits* referentes a transtornos mentais (ansiedade, depressão e suicídio) associados ao período pandêmico da covid-19. O segundo conjunto de dados também foi obtido a partir da plataforma Kaggle[6], constituído por postagens extraídas dos *subreddits r/depression* e *r/SuicideWatch*. Ao término do processo de tradução, foram agregados ao `DepreRedditBR` um total de 338.139 postagens.

Com a incorporação dos novos dados traduzidos, o conjunto de dados `DepreRedditBR` final resultou em um total de 541.573 de instâncias de textos.

Como inicialmente o foco deste trabalho era usar o conjunto de dados para o pré-treinamento do LLM e, tendo em vista que nem todas as fontes de dados possuiam os mesmos metadados estruturais, optou-se por gerar uma versão acomodando apenas a coluna com textos (títulos, postagens e comentários). Outra versão, considerando, além dos textos, os metadados de data, hora e alguma identificação que possa indicar perfis de usuários está em andamento.

3.2. Pré-processamento dos dados

Os dados brutos reunidos, após o término da etapa de coleta de dados, foram submetidos a rotinas típicas de pré-processamento de texto (Figura 1). As postagens do `DepreRedditBR`, então, foram submetidas à remoção de URLs (*Uniform Resource Locator*), emojis e emoticons. Haja vista que o corpo do texto de uma postagem no Reddit

[6]https://www.kaggle.com/datasets/xavrig/reddit-dataset-rdepression-and-rsuicidewatch

tem tamanho ilimitado, é comum encontrar muitas quebras de linha espalhadas ao longo do texto. Estas também foram removidas para evitar que a postagem seja fragmentada em várias partes. Ainda, foram removidas postagens com marcações especiais de controle $'NaN'$[7], $'[removed]'$ e $'[deleted]'$, os dois últimos referentes a fragmentos de texto removidos pelos moderadores do *subreddit* e substituídos pelas referidas marcações. Por fim, uma varredura para identificação de postagens duplicadas foi efetuada, com o objetivo de eliminar redundâncias ao introduzir os dados provenientes dos conjuntos de dados em inglês (traduzidos).

Ao término da etapa de pré-processamento, o conjunto de dados `DepreRedditBR` foi finalizado com um total de 509.675 mil instâncias, composta de títulos, postagens e comentários associados ao domínio da depressão. A Tabela 1 apresenta um extrato do `DepreRedditBR` após os tratamentos realizados.

Tabela 1. Amostra do conjunto de dados DepreRedditBR.

Texto
Planejando repetidamente acabar com minha vida, mas não consigo. Tudo o que faço sinto que falhei ou magoei as pessoas que amo. Sempre pareço sabotar algo quando tudo esta indo bem. Eu sei que meus dias estão contados. Eu não sei como escapar disso eu sou um fracasso
Não posso, já estou farto. Estou tão cansado que nada esta melhorando. eu tentei tanto, mas não consigo mais. É demais, muito cansativo. Meu último grupo de amigos estava certo. Eu não mereço estar vivo. Faz muito tempo que não me sinto real; faz mais tempo que não me sinto eu mesmo, só quero sentir que pertenço novamente. Deus, eu faria qualquer coisa por um abraço agora. Por que viver tem que ser tão cansativo? não posso mais fazer isso.
Meu psiquiatra me revelou meu diagnóstico há alguns dias. Sofro de depressão grave e tenho ansiedade. Estou tomando medicação e recentemente minhas doses aumentaram. Gostaria de falar com alguém em situação semelhante ou apenas conversar com alguém em geral...

4. Aplicabilidade do DepreRedditBR

O conjunto de dados produzido neste trabalho foi empregado até o momento em três frentes: (i) No pré-treinamento de um LLM especializado na temática Depressão; (ii) Na análise temporal de postagens no intercurso da pandemia de COVID-19 e (iii) Em uma análise de frequência de termos.

Com respeito ao item (i), conforme mencionado, a produção do conjunto de dados textuais `DepreRedditBR` foi uma necessidade iminente para o pré-treinamento de um LLM especializado na temática "depressão", no idioma português brasileiro, nomeado `DepreBertBR`. O `DepreBertBR`, a ser publicado nos anais do SBBD 2024, é baseado no BERT (*Bidirectional Encoder Representations for Transformers*) [Devlin et al. 2019], um LLM de código aberto criado por pesquisadores do Google, que utiliza a arquitetura *Transformer*[8].

O pré-treinamento do `DepreBertBR` percorre, basicamente, as etapas de tokenização e pré-treinamento do modelo em si. Primeiramente, um tokenizador é criado e pré-treinado para gerar o vocabulário utilizando os textos constantes no conjunto `DepreRedditBR`. Na tarefa de tokenização, o texto é dividido em unidades menores denominadas de *tokens*. Cada token representa uma palavra ou subpalavra dentro do vocabulário. Assim, as sentenças do `DepreRedditBR` são mapeadas para uma sequência de representações numéricas determinadas pelos IDs dos tokens presentes no vocabulário.

[7]Indicador de que em tal ponto havia um vídeo ou foto no lugar de um texto.

[8]Transformer é uma arquitetura de redes neurais que utiliza uma estrutura codificador-decodificador junto com um mecanismo de atenção para realizar tarefas de processamento de linguagem natural.

Essas representações númericas são utilizadas como entradas de dados para a etapa de pré-treinamento, juntamente com o vocabulário criado e o tokenizador.

Após a tokenização, inicia-se a etapa de pré-treinamento do `DepreBertBR`. No pré-treinamento o objetivo é treinar o modelo para aprender sobre um contexto (um domínio do problema em que está inserido) e o idioma padrão, considerando o conjunto de dados utilizado na etapa de tokenização. Alguns hiperparâmetros como, por exemplo, a taxa de aprendizado e estratégia de avaliação são configurados antes de inicializar o pré-treino. O resultado do pré-treinamento do `DepreBertBR` implica em um modelo singularizado na compreensão da linguagem natural a partir de conteúdos textuais do domínio da depressão. Assim, ele pode ser útil na realização de diferentes tarefas relacionadas ao processamento da linguagem em que foi treinado (e.g., classificação de texto, resposta a perguntas).

Em particular, o modelo `DepreBertBR` foi avaliado em uma tarefa de classificação de texto. Para isso, torna-se necessário realizar um procedimento denominado "ajuste fino"(*fine tunning*), o que corresponde a realizar um refinamento das habilidades aprendidas pelo modelo para a tarefa alvo, utilizando um conjunto de dados direcionado à tarefa específica. O ajuste fino do `DepreBertBR` foi efetuado a partir de um conjunto de dados criado por [Sampath and Durairaj 2022]. Este conjunto de dados é constituído por postagens extraídas do Reddit e rotuladas manualmente por especialistas de domínio em um dos seguintes graus de depressão: ausente, moderada ou grave.

Como ele é originalmente composto por postagens no idioma inglês, foi necessário realizar sua tradução para o idioma português brasileiro utilizando, também, a API do Google Translate. Uma vez traduzido, o conjunto de dados específico para a tarefa de classificação também teve de passar pelas mesmas rotinas de tratamento de dados realizadas durante o pré-processamento do `DepreRedditBR` (e.g., remoção de marcações de controle, URLs, emoticons, emojis). O conjunto de dados resultante apresentou um quantitativo de 10.230 postagens rotuladas.

Diferentemente do cenário convencional de classificação, em que um modelo preditivo é treinado usando um dos populares algoritmos de aprendizado do Estado da Arte (e.g., Árvore de Decisão, Naive Bayes), no ajuste fino do `DepreBertBR`, o próprio modelo é instanciado como um classificador e perpassa pelas etapas de treinamento e teste para associar os padrões de texto às classes do problema. A avaliação do modelo `DepreBertBR` obteve os seguintes resultados médios utilizando as tradicionais métricas de precisão (ρ), revocação (σ) e f-measure ($f1$): $\rho = 0,89$, $\sigma = 0,86$ e $f1 = 0,87$. Os resultados foram otimistas, considerando que o conjunto de dados `DepreRedditBR` utilizado para treinamento do modelo `DepreBertBR` se mostrou promissor como fonte de dados para tarefas relacionadas ao domínio da depressão, no idioma português brasileiro.

Para o cenário de análise referente ao item (ii), um fragmento do conjunto de dados que possuía informações sobre data e hora das postagens foi utilizado. O objetivo foi analisar postagens do Reddit no intercurso da Pandemia de COVID-19, considerando três períodos: antes, durante e após a pandemia. O estudo revelou a evolução das discussões sobre a temática de transtornos mentais, particularmente a depressão, tendo sido constatado um aumento significativo de conteúdo depressivo durante e após a pandemia, o que destaca o impacto duradouro da crise na saúde mental. Observou-se também um

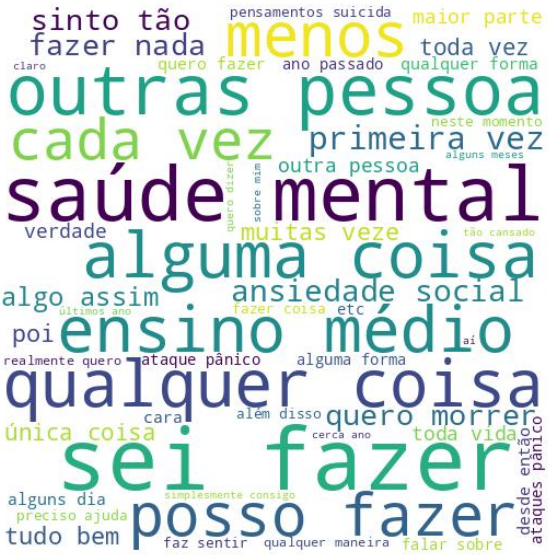

Figura 3. Nuvem de palavras

aumento nas postagens durante os dias úteis da semana durante a pandemia, enquanto que, nos períodos pré-pandemia e pós-pandemia, as postagens foram mais frequentes nos finais de semana. Além disso, constatou-se uma tendência de compartilhamento noturno de conteúdo depressivo em todos os períodos analisados, com a palavra "eu" sendo recorrentemente utilizada.Este trabalho está em processo de avaliação em um periódico [Estrela et al. 2024].

Adicionalmente, considerando o item (iii), foi realizada uma análise da frequência de ocorrência do vocabulário do `DepreRedditBR`, com o intuito de observar os termos que mais se destacam no contexto da depressão. A Figura 3 mostra uma nuvem de palavras gerada a partir da biblioteca *word_cloud*[9], recebendo como entrada todas as instâncias do conjunto de dados `DepreRedditBR`. Para esta análise, os dados passaram por tratamento básicos apenas de retirada de pontuações e de *stop words*. Analisando a nuvem de palavras, nota-se a proeminência de algumas expressões retiradas das postagem e inerentes ao contexto da depressão como, por exemplo, "tão cansado", "pensamentos suicidas", "ansiedade social", "ataques pânico"e "saúde mental". Tais termos puderam ser avaliados e ratificados pelos médicos especialistas como habituais nos diálogos estabelecidos pelos usuários com o transtorno em questão. As expressões destacadas ilustram o vocabulário contextualizado do conjunto de dados `DepreRedditBR` e seu potencial para outras análises.

5. Considerações Finais

A depressão é um transtorno mental que tem acometido milhões de pessoas no mundo todo, sendo motivo de grande preocupação pela OMS. A própria OMS estabeleceu a depressão como o mal do século, estimando que, até 2030, ela será a doença mais comum mundialmente. Com o advento das redes sociais, as pessoas passaram a compartilhar informações sobre sua vida e seu cotidiano, expondo sentimentos e emoções. Os usuários

[9]https://github.com/amueller/word_cloud

com tendências à depressão, muitas vezes, deixam transparecer por meio de palavras e frases indicativos de um estado depressivo [Cacheda et al. 2019]. Embora a literatura aponte iniciativas de pesquisa visando o reconhecimento de textos depressivos no idioma inglês, conjuntos de dados nessa temática no idioma português brasileiro ainda são escassos.

Este trabalho construiu um conjunto de dados textuais a partir de títulos, corpo do texto de postagens e de comentários com temática depressiva no idioma português brasileiro, obtidos a partir de publicações de usuários na rede social Reddit. O conjunto de dados, denominado `DepreRedditBR`, foi utilizado para o pré-treinamento de um LLM denominado `DepreBertBR`. O LLM pré-treinado foi ajustado para a tarefa de classificação de postagens de acordo com três possíveis graus de depressão. O conjunto de dados construído neste trabalho está disponível em https://zenodo.org/records/12761179.

Como trabalhos futuros, pretende-se incrementar o conjunto de dados com mais postagens relacionadas à depressão, possibilitando novos pré-treinamentos de modelos. Busca-se adicionalmente construir novos conjuntos de dados específicos para outros transtornos mentais como, por exemplo, o Transtorno de Ansiedade, Esquizofrenia, Transtornos alimentares, todos considerando o idioma português brasileiro. A ideia é possibilitar o desenvolvimento de aplicações que possam gerar alertas de sinais de transtornos mentais, levando à busca antecipada por ajuda médica.

Referências

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.

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

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.

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 Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, pages 128–137.

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.

Estrela, P., Andrade, L., Souza, D., Cunha, A., and Mendes, R. (2024). Análise de sentimentos em postagens do reddit no intercurso da pandemia de covid-19. *Submetido à Revista Principia*.

Herculano, A., Gomes, G., Souza, D., and Rêgo, A. (2022). Detecting signs of mental disorders on social networks: a systematic literature review. *DATA ANALYTICS 2022*, pages 55–61.

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.

Kristensen, C. H., Gomes, C. F. d. A., Justo, A. R., and Vieira, K. (2011). Normas brasileiras para o affective norms for english words. *Trends in Psychiatry and Psychotherapy*, 33:135–146.

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.

Nardi, A. E., da Silva, A. G., and Quevedo, J. (2021). *Tratado de Psiquiatria da Associação Brasileira de Psiquiatria*. Artmed Editora.

Pérez, A., Parapar, J., and Barreiro, Á. (2022). Automatic depression score estimation with word embedding models. *Artificial Intelligence in Medicine*, 132:102380.

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.

Sperling, O. V. and Ladeira, M. (2019). Mining twitter data for signs of depression in brazil. In *Anais do VII Symposium on Knowledge Discovery, Mining and Learning*, pages 25–32. SBC.

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.

Vedula, N. and Parthasarathy, S. (2017). Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health*, pages 127–136.

WHO (2023). World health organization. https://www.who.int/news-room/fact-sheets/detail/depression Last accessed 10 Julho 2024.