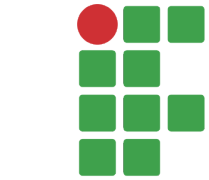


Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus João Pessoa

Programa de Pós-Graduação em Tecnologia da Informação

Nível Mestrado Profissional

AYRTON DOUGLAS RODRIGUES HERCULANO

DEPREBERTBR: UM MODELO DE LINGUAGEM PRÉ-TREINADO PARA O DOMÍNIO DA DEPRESSÃO NO IDIOMA PORTUGUÊS BRASILEIRO

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA

2024

AYRTON DOUGLAS RODRIGUES HERCULANO

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

Dissertação apresentada como requisito para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB.

Orientador: Prof. Dra. Damires Yluska de Souza Fernandes
Coorientador: Prof. Dr. Alex Sandro da Cunha Rego

João Pessoa

2024

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Nilo Peçanha - *Campus* João Pessoa, PB.

H539d Herculano, Ayrton Douglas Rodrigues.

DepreBERTBR : um modelo de linguagem pré-treinado para o domínio da depressão no idioma português brasileiro / Ayrton Douglas Rodrigues Herculano. – 2024.
72 f. : il.
Dissertação (Mestrado em Tecnologia da Informação) – Instituto Federal de Educação da Paraíba / Programa de Pós-Graduação em Tecnologia da Informação (PPGTI), 2024.
Orientação : Profa. D.ra Damires Yluska de Souza Fernandes.
Coorientação : Prof. D.r Alex Sandro da Cunha Rego.

1. Modelos de linguagem. 2. Transtorno mental - depressão. 3. Aprendizado por transferência. 4. BERT. 5. Reddit. I. Título.

CDU 004.43:613.8(043)

Bibliotecária responsável Lucrecia Camilo de Lima – CRB 15/132

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU*

MESTRADO PROFISSIONAL EM TECNOLOGIA DA INFORMAÇÃO

AYRTON DOUGLAS RODRIGUES HERCULANO

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

Dissertação apresentada como requisito para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós- Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB - Campus João Pessoa.

Aprovado em 31 de maio de 2024

Membros da Banca Examinadora:

Dra. Damires Yluska de Souza Fernandes

IFPB - PPGTI

Dr. Alex Sandro da Cunha Rego

IFPB

Dr. Francisco Dantas Nobre Neto

IFPB - PPGTI

Dr. Yuri De Almeida Malheiros Barbosa

UFPB

gov.br
Documento assinado digitalmente
YURI DE ALMEIDA MALHEIROS BARBOSA
Data: 30/07/2024 09:57:23-0300
Verifique em https://validar.iti.gov.br

João Pessoa/2024

Documento assinado eletronicamente por:

- **Damires Yluska de Souza Fernandes**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 05/06/2024 10:52:48.
- **Francisco Dantas Nobre Neto**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 05/06/2024 14:05:32.
- **Alex Sandro da Cunha Rego**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 05/06/2024 15:24:26.

Este documento foi emitido pelo SUAP em 24/05/2024. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 566593
Código de Autenticação: 3d0aee15f3

Dedico este trabalho aos meus pais, Herculano e Rosa, meus irmãos, Hércules e Wesley e minha esposa Renata. Você são a minha base e fortaleza.

AGRADECIMENTOS

Entrar em um mestrado profissional sempre foi um sonho e ao mesmo tempo, um desafio. Durante todo este percurso sempre fui desafiado a aprender e crescer, como pesquisador, como profissional e também como pessoa. Nem sempre é fácil conciliar tantos mundos, trabalho, família, pesquisas, mas a vontade se crescer sempre foi o combustível para adquirir mais conhecimento. Após esses anos, meu sentimento é de gratidão.

Primeiramente, gostaria de agradecer a Deus. Sem Ele em minha vida não teria conseguido superar tantos desafios ao longo dessa jornada. Em momentos difíceis foi a fé no Senhor que me sustentou firme e esperançoso para alcançar os objetivos. Até aqui o Senhor me ajudou.

Gostaria de agradecer aos meus pais e irmãos que compreenderam minha ausência em momentos com a família. Se cheguei até aqui é porque pude contar com o apoio deles.

À Renata, minha esposa, que foi uma pessoa fundamental durante toda a trajetória da pesquisa, suportando e me apoiando nessa caminhada. Muitas vezes, compreendendo a minha ausência e abdicando de momentos especiais em nossas vidas. Esse sonho também é seu.

Aos meus orientadores Damires e Alex, que tiveram intensa dedicação no desenvolvimento da nossa pesquisa. Não pouparam esforços para contribuir de várias formas com o nosso trabalho. A orientação de vocês me ajudou não só na pesquisa, mas também contribuíram no meu desenvolvimento pessoal e profissional.

A todos do grupo de pesquisa DepreTracker que em muitos momentos puderam me ajudar, foi um aprendizado mútuo, vocês contribuíram bastante e espero ter contribuído da mesma forma também.

À empresa E-ticons, a qual faço parte, que em muitos momentos soube compreender as minhas necessidades em me ausentar do trabalho, flexibilizando meus horários de trabalho para cumprir minhas atividades no mestrado.

Finalmente, agradeço ao Instituto Federal da Paraíba - campus João Pessoa, em especial aos professores do PPGTI e ao meus colegas de curso.

RESUMO

Transtornos mentais, tais como ansiedade e depressão, são caracterizados por causar distúrbios significativos na cognição e regulação emocional de uma pessoa. Particularmente, a depressão vem sendo alerta de preocupação pela Organização Mundial de Saúde diante de sua crescente expansão em escala mundial. Além de minar a autoestima do ser humano, em casos mais graves, esse transtorno pode levar à morte. As redes sociais vêm se tornando um espaço cada vez mais convidativo para que usuários com tendência depressiva possam expor seus sentimentos por meio de postagens. Pesquisas para detecção de indícios de depressão em redes sociais no idioma português brasileiro ainda são escassas, tendo em vista que a maioria dos trabalhos na literatura se concentra em soluções para o idioma inglês. Neste panorama, esta dissertação de mestrado propõe uma abordagem para classificar postagens de redes sociais no idioma português brasileiro com tendências depressivas. A abordagem apoia-se na construção de um modelo de linguagem pré-treinado denominado DepreBERTBR, baseado no modelo de linguagem BERT, focado no domínio da depressão para o idioma português brasileiro. Com o conhecimento adquirido durante o pré-treinamento, a partir de um corpus construído com postagens extraídas de subcomunidades do Reddit ligadas a transtornos mentais, o DepreBERTBR foi ajustado para a tarefa de classificação de textos considerando três graus de depressão: ausente, moderado ou grave. Como resultado, o DepreBERTBBR alcançou um valor médio de 0,87 para *F1-score*, considerando uma validação cruzada com 10 dobras, demonstrando que o modelo desenvolvido foi eficaz em classificar, de acordo com o grau de depressão, textos de postagens do Reddit com teor depressivo no idioma português do Brasil, em comparação a outros modelos de linguagem no idioma português do Brasil, mostrando-se ser bastante competitivo para tarefas de classificação e, em especial, para detectar indícios de depressão.

Palavras-chaves: Modelos de linguagem; Aprendizado por transferência; BERT; Classificação de textos; Depressão; Reddit.

ABSTRACT

Mental disorders, such as anxiety and depression, are characterized by causing significant disturbances in a person's cognition and emotional regulation. In particular, depressive disorder has been one of the main priority conditions covered by World Health Organization (WHO), given its increasing ocurrences expansion on a global scale. In addition to undermining human self-esteem, in more severe cases, it may lead humans to death. In the light of social networks, users who tend to have depression may find some of them inviting to express feelings, by means of posts. Research to detect signs of depression on social networks in the Brazilian Portuguese language is still scarce, considering that the majority of available works in the literature focus on solutions for the English language. In this scenario, this master's thesis proposes an approach to classify social media posts in Brazilian Portuguese with depressive tendencies. The approach is based on the construction of a pre-trained language model, called DepreBERTBR, based on the BERT language model, focused on the domain of depression for the Brazilian Portuguese language. With the knowledge acquired during pre-training,from a given corpus built with posts extracted from some Reddit subcommunities linked to mental disorders, DepreBERTBR has been tuned for a text classification task underlying three possible degrees of depression, as follows: absent, moderate or severe. As a result, DepreBERTBBR achieved an average value of 0.87 for F1-score, considering a 10-fold cross-validation, demonstrating that the model developed was effective in classifying texts from Reddit posts with kinds of depressive contents in Brazilian Portuguese, with respect to other large language models in Brazilian Portuguese. Thus, it shows to be competitive for classification tasks, particularly, in the light of detecting signs of depression given portuguese texts.

Keywords: Language models; Transfer learning; BERT; Text classification; Depression; Reddit.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM: Aprendizado de Máquina

AMS: Aprendizado de Máquina Supervisionado

AMNS: Aprendizado de Máquina Não-Supervisionado

AMSS: Aprendizado de Máquina Semi-Supervisionado

AP: Aprendizado Profundo

API: Application Programming Interface

AS: Análise de Sentimentos

AT: Aprendizado por Transferência

BERT: Bidireciton Enconder Representations from Transformers

BoW: bag-of-words

CBOW: Continuos Bag of Words

CNN: Convolutional Neural Networks

IDB: Inventário de Depressão de Beck

KNN: K-nearest neighbors

LLM: Large Language Model

LSTM: Long short term memory

Masked LM: Masked Language Modeling

mBERT: Multilingual BERT

ML: Modelos de Linguagem

MLP: Multilayer perceptron

NSP: Next Sentence Prediction

OMS: Organização Mundial de Saúde.

PLN: Processamento de Linguagem Natural

RN: Redes Neurais

RNA: Redes Neurais Artificiais

RNN: Redes Neurais Recorrentes

SVM: Support Vector Machine

TEPT: Transtorno de Estresse Pós-Traumático

TF: Term Frequency

TF-IDF: Frequência do Termo - Inverso da Frequência no Documento

URLs: Uniform Resource Locator

SUMÁRIO

1 INTRODUÇÃO

Este capítulo introduz o contexto em que a presente dissertação encontra-se inserida. Para tal, são apresentadas a motivação, justificativa, questões de pesquisa, objetivos e a estrutura do documento.

1.1 Motivação e Definição do Problema

De acordo com o American Psychiatric Association (2013), um transtorno mental é formalmente definido como uma síndrome caracterizada por um distúrbio clinicamente significativo na cognição de um indivíduo, regulação emocional ou comportamento que reflete uma disfunção nos processos psicológicos, biológicos ou de desenvolvimento subjacentes ao funcionamento mental. Exemplos de transtornos mentais mais comuns incluem (RÍSSOLA; LOSADA; CRESTANI, 2021): transtorno de ansiedade, transtorno alimentar (e.g., anorexia, bulimia), automutilação, afetivo bipolar, Transtorno de Estresse Pós-Traumático (TEPT), esquizofrenia e depressão.

Dentre os transtornos mentais citados, a depressão é vista como o Mal do Século pela Organização Mundial de Saúde (OMS), que em 2022 estimou um número de cerca de 280 milhões de pessoas acometidas por esse distúrbio em todo o mundo (WHO, 2023). Neste mesmo período, no Brasil, cerca de 5,7% da população sofria com a depressão, sendo este índice o maior entre os países latino-americanos. Em comparação aos países das três Américas, o Brasil ficou atrás somente dos Estados Unidos, onde aproximadamente 5,9 % da população apresentaram sinais de depressão (ORGANIZATION, 2017). A previsão atual da OMS indica que até 2030, a depressão será a doença mais comum em todo o mundo.

A depressão pode atingir qualquer pessoa, independentemente do sexo ou idade, devido a fatores de ordem genética (e.g., histórico familiar de depressão, doenças crônicas), bioquímicos (e.g., uso abusivo de drogas lícitas ou ilícitas), psicológicos (e.g., luto por perda afetiva, baixa autoestima, estresse) e sóciofamiliares (e.g., isolamento social, situação financeira desfavorável, discriminação) (CUNHA; BASTOS; DUCA, 2012). Reconhecer pessoas com depressão não é uma tarefa fácil, haja vista que há o receio delas esconderem seus sintomas por medo de serem julgadas/rejeitadas ou até mesmo fingirem que está tudo bem, se mostrando felizes para outras pessoas, mas internamente se sentindo tristes ou desanimadas. O diagnóstico precoce pode ajudar significativamente no tratamento e na cura da doença.

De acordo com Cacheda et al. (2019) e Vedula e Parthasarathy (2017), a prevenção contra a depressão pode acontecer de diferentes formas, uma delas por meio do monitoramento de comportamento de usuários em redes sociais. Estudos realizados com jovens adultos nos Estados Unidos mostram que há um crescente interesse no uso de redes sociais em busca de um

bem-estar psicológico (LIN et al., 2016).

As redes sociais tornaram-se um espaço virtual muito popular entre a maioria das pessoas, sendo utilizadas para várias finalidades como, por exemplo, divulgar serviços e produtos (compra e venda), compartilhar e buscar informações sobre temas variados, estabelecer contatos pessoais, procurar emprego ou preencher vagas, entre outros. 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.

As postagens compartilhadas publicamente pelos usuários de redes sociais são ricas em informações que podem trazer embutidas particularidades relacionadas a interesses pessoais, comportamento, viés político, opinião ou sentimentos (DUQUE; RAYMUNDO; NETO, 2018). As interações realizadas pelos usuários evidenciam uma linguagem que pode denotar a presença de emoções e sentimentos dos mais variados, os quais podem indicar percepções de inutilidade, culpa, solidão, abandono e ódio a si próprio, podendo assim caracterizar indício de algum tipo de transtorno mental, como a depressão (CHOUDHURY et al., 2013).

Usuários acometidos com algum transtorno mental, como a depressão, tendem a demonstrar comportamento online distinto de outros usuários que não possuem nenhum transtorno perceptível (SPERLING; LADEIRA, 2019; CHOUDHURY et al., 2013). Quando interagem em espaços virtuais focados no compartilhamento ou desabafos emocionais dessa natureza (e.g., comunidades online sobre o tema "depressão"), usuários se sentem mais à vontade para demonstrar seus sentimentos ou angústias, em busca de apoio ou de indentificar-se com alguém (UBAN; CHULVI; ROSSO, 2021). Entretanto, detectar depressão é uma tarefa complexa, visto que a identificação desse transtorno decorre da combinação de sinais e sintomas que possam persistir diariamente por, no mínimo, duas semanas. Alguns sintomas também podem estar associados a outros transtornos mentais (NARDI; SILVA; QUEVEDO, 2021). Sendo assim, é necessário que as pessoas acometidas pela doença tomem a iniciativa de buscar ajuda médica. Às vezes, o diagnóstico e posterior tratamento pode acontecer de forma tardia diante da evolução dos sintomas já desenvolvidos e alterações no estado psicológico do paciente.

Pesquisas nas áreas de Psiquiatria, Psicologia, Sociolinguística e Neurociência (VEDULA; PARTHASARATHY, 2017; ROSA et al., 2018; CACHEDA et al., 2019), associadas a técnicas computacionais específicas como, por exemplo, a análise de sentimentos, buscam aprimorar a compreensão da relação entre o comportamento das pessoas, seus sentimentos e emoções, utilizando como fonte de dados os textos postados em redes sociais.

A Análise de Sentimentos (AS) é uma das áreas da Computação que pode processar dados textuais oriundos de postagens de redes sociais, com o objetivo de identificar opiniões, sentimentos ou posicionamentos do público envolvido (DENG; SINHA; ZHAO, 2017). As técnicas de AS são capazes de identificar se uma publicação apresenta conotação positiva, negativa ou neutra referente ao conteúdo postado, de maneira a perceber o julgamento do usuário sobre

determinado assunto (GOVINDASAMY; PALANICHAMY, 2021). A AS está intimamente relacionada ao uso de estratégias de Processamento de Linguagem Natural (PLN), Aprendizado de Máquina (AM) e Aprendizado Profundo (AP) (OLIVEIRA et al., 2022). No âmbito do AP, modelos de linguagem computacionais gerados por redes neurais têm sido cada vez mais empregados em pesquisas e ferramentas associadas à análise de sentimentos e mineração de textos. Modelos de Linguagem (ML) também têm sido buscados para apoiar tarefas de detecção de sinais de depressão a partir de postagens em redes sociais (JI et al., 2022; POŚWIATA; PEREŁKIEWICZ, 2022). Nota-se, entretanto, que a grande maioria desses trabalhos constroi ou usa modelos de linguagem e implementa tarefas de PLN e/ou classificação de textos considerando o idioma inglês.

Há modelos como o BERT (*Bidireciton Enconder Representations from Transformers*) que foram desenvolvidos e treinados com dados no idioma inglês (DEVLIN et al., 2019). Apesar de existirem modelos de linguagem multilingue como, por exemplo, o mBERT(*multilingual* BERT), estudos como o trabalho de Martin et al. (2020) para o idioma francês, Cañete et al. (2023) para o idioma espanhol e Polignano et al. (2019) para o idioma italiano, demonstram que utilizar um modelo pré-treinado monolíngue pode ser mais eficiente para tarefas de PLN.

As pesquisas desenvolvidas por Lee et al. (2020) no domínio biomédico e Alsentzer et al. (2019) para o domínio clínico, ambos com dados no idioma inglês, mostram que ML pré-treinados para um domínio específico têm se destacado em várias tarefas de PLN (JI et al., 2022).

Considerando o cenário exposto, algumas questões de pesquisa norteiam este trabalho de pesquisa, a saber:

- QP1: Como classificar postagens de redes sociais no idioma português do Brasil considerando um possível grau de depressão?
- QP2: Utilizar um modelo de linguagem pré-treinado no domínio específico da depressão no idioma português do Brasil pode ajudar a determinar o grau de depressão percebido em postagens de redes sociais?

1.2 Objetivos

Com o propósito de responder às questões de pesquisa apresentadas neste capítulo, este trabalho propõe a criação de um *corpus* na língua portuguesa brasileira com postagens de teor depressivo, o qual é usado para fins de pré-treinamento de um modelo de linguagem baseado no BERT. O modelo gerado é ajustado para a tarefa de classificação de texto de postagens no Reddit, conforme os seguintes graus de depressão: ausente, moderada ou grave. Mais precisamente, a seguir são apresentados o objetivo geral e os objetivos específicos da dissertação.

1.2.1 Objetivo geral

Desenvolver uma abordagem de PLN baseada em um modelo de linguagem pré-treinado que, ajustado para a tarefa de classificação de textos, possa classificar postagens de redes sociais no idioma português do Brasil de acordo com um grau de depressão.

1.2.2 Objetivos específicos

- Construir um *corpus* com textos de postagens de teor depressivo no idioma português brasileiro, extraído a partir da rede social Reddit;
- Pré-treinar um modelo de linguagem a partir dos dados do *corpus* criado utilizando o modelo de linguagem pré-treinado BERT;
- Realizar o *fine tuning* (ajuste fino) para o modelo de linguagem pré-treinado para classificar postagens de acordo com níveis de depressão.
- Avaliar o desempenho do modelo pré-treinado desenvolvido, comparando-o com outros modelos de linguagem de domínio geral e de domínio geral com parte dos dados relativos à depressão e avaliá-lo com respeito à tarefa de classificação de postagens com possível teor depressivo.

1.3 Estrutura do Documento

Os capítulos subsequentes estão organizados da seguinte maneira: o Capítulo 2 introduz o referencial teórico acerca dos principais conceitos cobertos nesta dissertação. O Capítulo 3 descreve alguns trabalhos relacionados. O Capítulo 4 apresenta a abordagem baseada na construção do modelo de linguagem DepreBERTBr. O Capítulo 5 descreve a avaliação experimental realizada e os resultados obtidos. Finalmente, o Capítulo 6 apresenta considerações finais acerca do trabalho e propostas de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo discorre sobre os principais conceitos abordados ao longo deste trabalho, começando com concepções sobre o transtorno da depressão associado à utilização de redes sociais. Em seguida, descreve noções sobre aprendizado de máquina e aprendizado por transferência. O Capítulo também introduz uma explanação sobre análise de sentimentos e processamento de linguagem natural. Por fim, são apresentados conceitos sobre modelos de linguagem, arquitetura *transformers* e o modelo de linguagem BERT.

2.1 Transtorno da Depressão

A depressão é uma doença cada vez mais evidente em nossa sociedade. Trata-se de um dos transtornos mentais mais crescentes em todo o mundo e uma das principais doenças responsáveis por tentativas de suicídio (RÍSSOLA; LOSADA; CRESTANI, 2021). A depressão é uma perturbação mental que atinge o lado emocional do indivíduo, causando-lhe, muitas vezes, baixa autoestima, sentimentos de inutilidade, falta de ânimo, tristeza, sentimentos de culpa e, ainda, desejo de morte em situações mais graves (WHO, 2022). Fadiga, cansaço, dores crônicas, insônia e ansiedade são também outros sintomas apresentados por pessoas que têm depressão (NARDI; SILVA; QUEVEDO, 2021). A diminuição de energia, o desânimo, a falta de vontade e de iniciativa acontencem em intensidades variáveis, categorizando o grau de depressão em leve, moderada ou grave (MIGUEL et al., 2021).

Conforme o Manual Diagnóstico e Estatístico de Transtornos Mentais (DSM-5), o transtorno depressivo é classificado em subtipos, a saber (American Psychiatric Association, 2013):

- **Transtorno disruptivo de desregulação do humor**: adequação em que o indivíduo apresenta sempre humor irritável e frequentes crises de raiva;
- **Transtorno depressivo maior:** situação em que os sintomas depressivos diários permanecem por no mínimo duas semanas, causando malefícios, sendo este o tipo predominante na população em geral;
- **Transtorno depressivo persistente:** os sintomas depressivos perduram por no mínimo dois anos. Anteriormente, este subtipo da depressão era chamado de distimia;
- **Transtorno disfórico pré-menstrual:** alterações de humor e ocorrência de sintomas como ansiedade, irritabilidade, alterações na libido, sono e apetite durante o período que antecede à menstruação na maioria dos ciclos menstruais, em um período de um ano;

- **Transtorno depressivo induzido por substâncias:** Este subtipo do transtorno depressivo está relacionado ao uso de substâncias e/ou medicamentos. Apresenta sintomas passageiros de intoxicação;
- **Transtornos depressivos devido a outras condições médicas gerais:** Apresenta um cenário de depressão relacionado a doenças clínicas.

A heterogeneidade clínica, ou seja, a combinação de sintomas, sinais e critérios de realização de diagnósticos que podem se manifestar simultaneamente, são fatores que tendem a dificultar tratamentos mais específicos do transtorno depressivo (NARDI; SILVA; QUEVEDO, 2021). Para diagnosticar indivíduos com depressão é necessário realizar o atendimento inicial por um profissional especializado, partindo da premissa de que a pessoa tomou a iniciativa em buscar auxílio médico. Profissionais da área de saúde mental utilizam sistematicamente questionários como, por exemplo, o Inventário de Depressão de Beck (IDB), para auxiliar no diagnóstico de um paciente com ou sem depressão (GORENSTEIN; ANDRADE, 1998). O referido questionário é composto por 21 questões associadas a aspectos emocionais e comportamentais de um indivíduo, como, por exemplo, tristeza, pessimismo e pensamentos suicidas (TLELO-COYOTECATL; ESCALANTE; GÓMEZ, 2022). Cada questão possui 4 opções de respostas, podendo atingir de 0 a 3 pontos. De acordo com a somatório total dos pontos obtidos ao longo das 21 questões, o indivíduo pode ser classificado em um dos níveis de depressão: leve, moderada, grave ou ausente (GORENSTEIN; ANDRADE, 1998).

Apesar dos meios de avaliação utilizados por especialistas, muitas vezes, antes de receberem diagnóstico clínico, as pessoas podem demonstrar indícios de depressão através comportamentos e linguagens, tanto no mundo real quanto em meios digitais como as redes sociais (CHOUDHURY et al., 2013).

2.2 Reddit

O crescimento do uso de redes sociais modificou a rotina das pessoas. Os usuários passaram a interagir com seus seguidores e a compartilhar seu cotidiano, sentimentos, humor e emoções que estão passando. Isso tem possibilitado a condução de análises de dados e de percepção de comportamentos de usários nas redes sociais (LIN et al., 2017). Pesquisas recentes têm utilizado dados da rede social Reddit[1] para realizar experimentos de detecção de sintomas relacionados à depressão, utilizando postagens e comentários realizados pelos usuários (HERCULANO et al., 2022).

O Reddit é uma plataforma de rede social que permite o compartilhamento de conteúdo em formato de texto, imagem e links (PÉREZ; PARAPAR; BARREIRO, 2022). O Reddit possui comunidades, também chamadas de *subreddit*, um espaço onde os usuários com o mesmo

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

interesse postam sobre um assunto específico (JI et al., 2022). Exemplos de subreddits incluem `r/depression`, `r/SuicideWatch`, `r/Anxiety`, `r/AnsiedadeDepressao` e `r/Desabafos`. De forma geral, uma postagem no Reddit possui um título, um texto referente à postagem e um subreddit em que a postagem está associada, como pode ser observada na Figura 1. Uma postagem também pode conter comentários relacionados. Os textos postados no Reddit não possuem um limite fixo de caracteres, isto é, podem apresentar um conteúdo textual curto ou extenso. Para o título e comentários, há uma limitação de 300 caracteres e 10.000 caracteres, respectivamente.

Figura 1 – Postagem no Reddit.

Fonte: https://www.reddit.com/r/AnsiedadeDepressao/ .

A Figura 1 mostra um exemplo de postagem no Reddit. Para preservar a privacidade dos usuários, seus nomes foram substituídos por "usuário anonimizado". Na parte superior da figura, pode-se observar o subreddit "r/AnsiedadeDepressao"em que foi realizada esta postagem. O título da postagem aparece logo em seguida: "Depressão e Ansiedade". Abaixo do título é mostrado o conteúdo da postagem associado ao principal tópico do subreddit. Na parte inferior encontram-se os comentários de outros usuários desta comunidade. Percebe-se que o texto da postagem mostra o desabafo e relato de alguém que foi diagnosticado com depressão. O autor da postagem expõe detalhes sobre seu comportamento, seus sentimentos de culpa e triteza, dessa forma, demonstrando sinais de depressão. O comentário feito por outro usuário evidencia os próprios sentimentos, ao mesmo tempo em que demonstra apoio ao autor da postagem, mostrando que a interação entre os usuários do reddit pode produzir dados reais que podem ser utilizados por soluções computacionais para fins específicos.

A popularização do uso de redes sociais têm contribuído para surgimento de diferentes soluções computacionais que utilizam técnicas de análise de sentimentos, processamento de linguagem natural, aprendizado de máquina e modelos de linguagem grande com a finalidade de detectar depressão (POŚWIATA; PEREŁKIEWICZ, 2022; JI et al., 2022). Modelos de linguagem grande (*Large Language Models*) (LLM) são abordados com mais detalhes na Seção 2.6.4, mas, de modo geral, LLMs são modelos baseados em aprendizado profundo que são pré-treinados com uma enorme quantidade de dados textuais para aprenderem sobre o contexto das palavras no texto (PAES; VIANNA; RODRIGUES, 2023).

2.3 Aprendizado de máquina

A habilidade de aprendizagem é uma aptidão dos seres humanos que permite aprimorar-se à medida que eles realizam tarefas semelhantes. De forma análoga, sistemas computacionais foram desenvolvidos para aprender similarmente ao aprendizado humano, ou seja, baseado em experiências. Essa abordagem é conhecida como Aprendizado de Máquina (AM). O AM é uma subárea da Inteligência Artificial (IA) que provê às máquinas computacionais a habilidade de obter conhecimento de forma automática, a partir de dados e experiências passadas, sem que sejam explicitamente programado (MONARD; BARANAUSKAS, 2003). Algoritmos de AM fazem uso de uma perspectiva indutiva para produzir conhecimentos novos e predizer situações futuras (MITCHELL et al., 1997). Para Ligthart, Catal e Tekinerdogan (2021), de modo geral, as abordagens de aprendizado de máquina podem ser divididas em três categorias: Aprendizado Supervisionado, Aprendizado Não Supervisionado ou Aprendizado Semi-Supervisionado. Ainda, inserida no campo do AM, há o Aprendizado Profundo e o Aprendizado por transferência, os quais serão brevemente conceituados adiante.

O Aprendizado de Máquina Supervisionado (AMS) é um tipo de aprendizado de máquina em que as instâncias de dados são categorizadas através de rótulos. Esses dados são utilizados como entrada para treinar um modelo preditivo, de maneira que o modelo possa aprender relações entre instâncias de mesmo rótulo e seja capaz de realizar previsões de rótulos de instâncias de dados posteriomente desconhecidos (LIGTHART; CATAL; TEKINERDOGAN, 2021). O AMS pode ser implementado como um problema de classificação ou regressão (HARRINGTON, 2012). Na classificação, o modelo analisa variáveis de um conjunto de dados de entrada com o objetivo de determinar uma variável de saída (e.g. verdadeiro ou falso, depressivo ou não depressivo). Para a tarefa de regressão, a variável de saída é um valor contínuo, geralmente, pertencendo ao conjunto dos números reais. (ALPAYDIN, 2020). Alguns exemplos de algoritmos típicos de tarefas de Aprendizado Supervisionado são o *K-nearest neighbors* (KNN), Árvore de Decisão, *Support Vector Machine* (SVM), *naive bayes*, regressão logística e regressão linear (HAN; KAMBER; PEI, 2012).

Na categoria de Aprendizado de Máquina Não Supervisionado (AMNS) recebe como entrada um conjunto de dados que não possui rótulos pré-definidos com o objetivo de identificar

similaridades ou relacionamentos entre os dados de entrada (ALPAYDIN, 2020). Posto isto, um algoritmo é configurado para analisar o relacionamento entre os exemplos de dados e, de acordo com o padrão encontrado, agrupar os exemplos de dados semelhantes em unidades distintas denominadas "clusters". Outro exemplo de tarefa do aprendizado não supervisionado são as regras de associação. O objetivo, neste caso, é encontrar uma relação entre as instâncias em um conjuntos dados e verificar, por exemplo, a frequência em que um grupo de instâncias ocorrem juntas (HARRINGTON, 2012). Exemplos de algoritmos de aprendizado não supervisionado são o K-means e o Apriori.

O AMSS é aplicado em cenários onde, geralmente, existe uma grande quantidade de dados não rotulados e poucos dados rotulados (HAN; KAMBER; PEI, 2012). Essa abordagem de aprendizado pode utilizar o autotreinamento ou o cotreinamento no desenvolvimento do classificador. Na abordagem de autotreinamento é construído um classificador usando os dados rotulados, e esse classificador tenta classificar os dados não rotulados. Já no cotreinamento dois ou mais classificadores são utilizados, cada classificador é treinado com uma parte dos dados, em seguida, cada classificador um ensina ao outro (HAN; KAMBER; PEI, 2012).

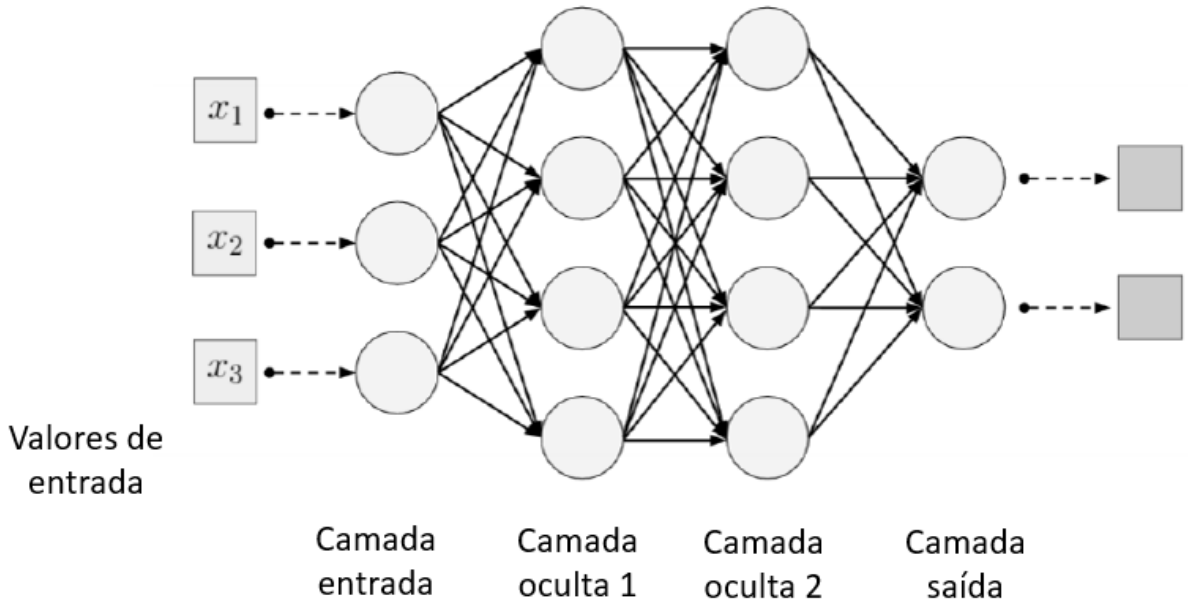
2.3.1 Aprendizado profundo

O aprendizado profundo (AP), em inglês, *deep leaning*, é uma subárea do AM capaz de processar grandes quantidades de dados buscando identificar e aprender os padrões e relacionamentos que esses dados apresentam através de camadas ocultas da rede neural (RN) (TAULLI, 2020). As RN são sistemas computacionais baseados na estrutura e funcionamento de um cérebro humano, sendo composto por unidades básicas chamadas de neurônios artificiais e organizados em camadas (GOODFELLOW; BENGIO; COURVILLE, 2016). Uma rede neural artificial (RNA) como a *Multilayer perceptron* (MLP) é formada por estruturas com várias camadas interconectadas de *perceptrons*, uma representação matemática que faz analogia a um neurônio biológico. Este tipo de RN é formado por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída (SILVA; VIEIRA, 2022). Redes MLP são chamadas de *feedforward networks* em razão da informação ser processada da camada de entrada até a camada de saída sem que seja necessária a retroalimentação entre as camadas ou unidades (GOODFELLOW; BENGIO; COURVILLE, 2016). A Figura 2 mostra a arquitetura de uma rede MLP com duas camadas ocultas. Os dados *x1, x2, x3* são passados para a camada de entrada da rede. As camadas internas (camadas ocultas), ou seja, que não são de entradas ou saída, processam os dados gerando a informação para a camada de saída (SILVA; VIEIRA, 2022).

Quando uma RNA, como a MLP, possui muitas camadas ocultas, são chamadas de Redes Neurais Profundas (RNP) (GOODFELLOW; BENGIO; COURVILLE, 2016). Além das MLPs, as Redes Neurais Recorrentes (*Recurrent Neural Networks* - RNN) e as e Redes Neurais Convolucionais (Convolutional Neural Networks - CNNs) são exemplos de RNP (SILVA; VIEIRA, 2022). As RNP têm a capacidade de aprender vários níveis de abstração dos dados,

permitindo produzir representações vetoriais que capturam semelhanças linguísticas, como, por exemplo, palavras presentes em um determinado texto (OLIVEIRA et al., 2022).

Figura 2 – Exemplo de Rede MLP com duas camadas ocultas.



Fonte:Adaptado de Silva e Vieira (2022)

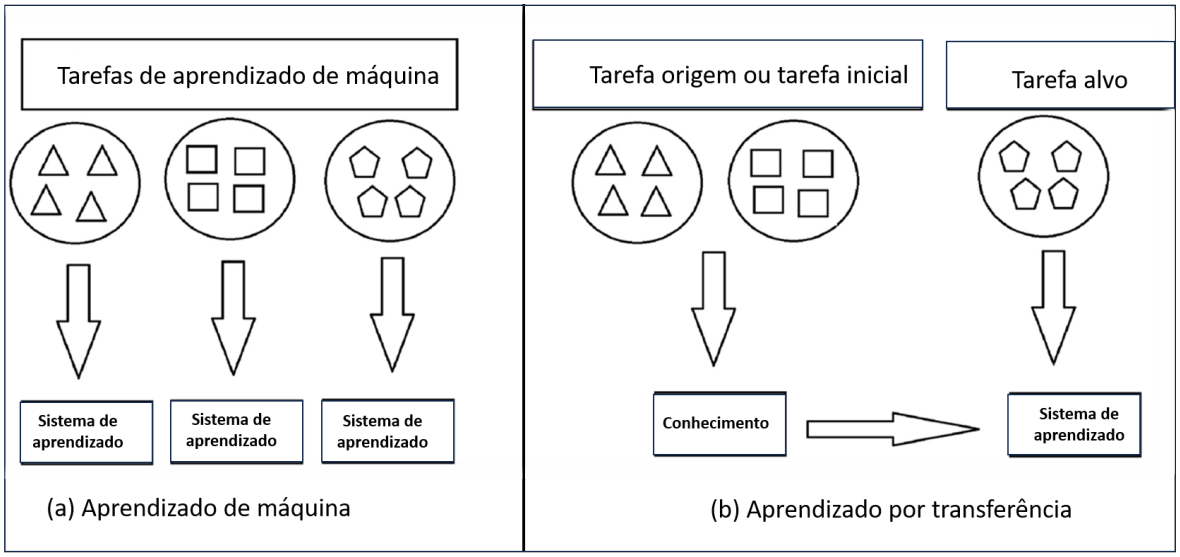
2.3.2 Aprendizado por transferência

O aprendizado por transferência (AT) é um técnica de AM em que um modelo base, geralmente um modelo de linguagem ou modelo de RNP, é treinado utilizando um conjunto de dados para uma tarefa inicial, sendo ajustado posteriormente para uma tarefa alvo. O conhecimento e características aprendidos pelo modelo base são transferidos para um segundo modelo ou modelo alvo, que é ajustado para uma tarefa alvo, normalmente associada a um domínio específico como, por exemplo, depressão. (YOSINSKI et al., 2014). Dessa forma, o modelo alvo tem seus pesos inicializados com os valores dos pesos do modelo base (SOUZA; NOGUEIRA; LOTUFO, 2020). Normalmente, o modelo base é treinado com grandes conjuntos de dados com o objetivo de ensiná-lo sobre as características desses dados. Em seguida, o modelo base pode ser adaptado a uma nova tarefa sem que seja necessário treiná-lo novamente desde o início (TUNSTALL; WERRA; WOLF, 2022).

Um exemplo de AT é o pré-treinamento de um modelo de linguagem que utiliza um *corpus* com milhões de sentenças de domínio geral (e.g. política, saúde, esporte, educação, entretenimento). Após aprender as características da linguagem como contexto, gramática e idioma, esse modelo base pode ser ajustado para realizar uma classificação de sentimentos

como, por exemplo, no domínio da depressão, utilizando um conjunto de dados rotulados menor. A Figura 3 mostra um comparativo entre métodos de AM e AT. No quadro (a) da figura são mostrados exemplos de como os métodos de AM aprendem sobre os dados apenas para um domínio específico, cada círculo representa um domínio de dados. Em contrapartida, o quadro (b) mostra que no método de AT, o modelo base aprende sobre vários domínios em uma tarefa inicial e, em seguida, transfere o conhecimento adquirido para uma tarefa alvo.

Figura 3 – Comparativo entre AM e AT.



Fonte: Adaptado de Pan e Yang (2009)

O AT busca utilizar modelos pré-existentes para resolver novas tarefas, sem a necessidade de desenvolver uma solução desde o princípio. Assim, melhorando o aprendizado a partir do conhecimento adquirido no pré-treinamento e reduzindo o tempo na implementação de um novo modelo (OLIVAS et al., 2009).

2.4 Análise de Sentimentos

A Análise de Sentimentos (AS) é um processo que se apoia na utilização de métodos computacionais com o propósito de extrair opiniões, sentimentos e emoções em linguagem natural, de forma automática (BENEVENUTO; RIBEIRO; ARAÚJO, 2015). Normalmente, muitos problemas no contexto da AS lidam com abordagens direcionadas à análise de texto (foco deste trabalho de mestrado), porém é possível encontrar trabalhos de extração de AS a partir de conteúdo de áudio, vídeo ou imagens (RAO et al., 2021).

A AS também é chamada de Mineração de Opinião, pois realiza a identificação, análise e classificação de opinião expressas em um texto (TARDELLI; DIAS; FRANÇA, 2019). A AS é apontada como uma área multidisciplinar por incluir conhecimentos provenientes dos domínios da Psicologia, Sociologia, Processamento de Linguagem Natural (PLN) e AM (LIGTHART;

CATAL; TEKINERDOGAN, 2021). Alguns conceitos e termos pertinentes ao contexto da AS são resumidamente descritos a seguir (BENEVENUTO; RIBEIRO; ARAÚJO, 2015):

- **Polaridade:** indica o grau de positividade ou negatividade de um texto. Alguns trabalhos tratam a polaridade de forma binária (positivo ou negativo) ou de modo ternário (positivo, negativo ou neutro). Um exemplo positivo é a frase Nossa, o dia hoje está realmente lindo, enquanto que a frase Hoje está tudo nublado e sem vida remete a uma polaridade negativa. Por outro lado, uma frase como Hoje é 19 de novembro não possui polaridade, logo, denota uma classificação de neutralidade;

- **Força do sentimento:** caracteriza a intensidade de um sentimento ou da polaridade. Normalmente é um ponto flutuante no intervalo de -1 a 1, muitas vezes sendo necessário o uso de um *threshold* (limiar) para identificar a neutralidade de uma sentença. Uma notícia boa ou ruim pode apresentar a força do sentimento como, por exemplo, na frase "A seleção brasileira venceu com grande vantagem a copa do mundo!". Nesse exemplo, para os torcedores do Brasil, a notícia pode representar um sentimento bom ou uma polaridade positiva;

- **Sentimento/emoção:** descreve um sentimento específico presente em uma mensagem (e.g., raiva, surpresa, alegria, etc). A frase "Nossa, meu artigo foi escolhido como o melhor da conferência!"pode indicar sentimentos de supresa e felicidade;

- **Subjetividade vs. objetividade:** uma sentença objetiva possui normalmente um fato ou uma informação, enquanto sentenças subjetivas expressam sentimentos pessoais e opiniões. Textos informais, como aqueles coletados de redes sociais, tendem a ser mais subjetivos. Um exemplo para uma sentença objetiva é a frase "Hoje o dia está ensolarado e muito quente". Já a frase "Estou triste e sem energia para viver"contempla um aspecto de subjetividade na sentença, haja visto que ela expressa um sentimento particular do autor da sentença.

  Como exemplo de AS, no contexto da depressão, uma publicação poderia ser classificada como depressiva ou não depressiva. Também, poderia seguir uma classificação conforme pontuação de severidade a partir de referências da Saúde como o IDB.

2.5 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área de pesquisa relacionada à Inteligência Artificial com o objetivo de pesquisar e propor técnicas para o processamento da linguagem humana por sistemas computacionais (CASELI; NUNES, 2023). O PLN envolve um conjuntos de técnicas para análise de texto com o objetivo de fazer com que aplicações computacionais compreendam o seu significado similarmente à forma como os seres humanos fazem (LIDDY, 2001).

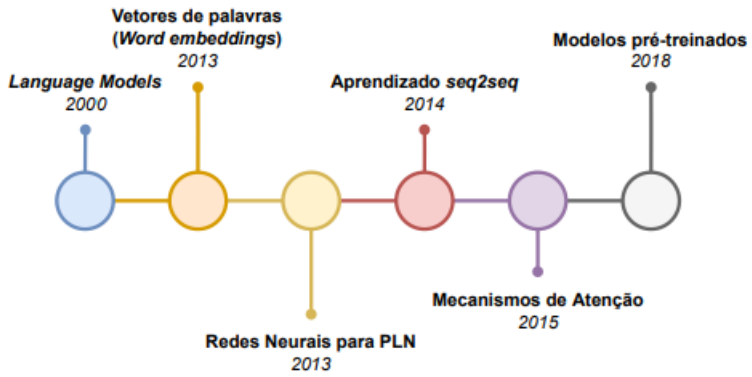
Atualmente, o PLN utiliza técnicas de AP com o propósito de entender melhor uma linguagem, o contexto e significado das palavras para solucionar problemas, ou tarefas, como, por exemplo, Classificação de Textos, Reconhecimento de Entidade Nomeada, Modelagem de Linguagem e Sumarização (OLIVEIRA et al., 2022).

A tarefa de Classificação de Texto tem como finalidade classificar se um texto como, por exemplo, uma postagem em rede social apresenta ou não indícios de depressão (OLIVEIRA et al., 2022). O Reconhecimento de Entidade Nomeada busca identificar no texto elementos (palavras) que se referem a algo do mundo real, como, por exemplo, uma Pessoa, Local, Data, Objeto e Organização (TUNSTALL; WERRA; WOLF, 2022). A tarefa de Modelagem de Linguagem busca identificar qual a próxima palavra em uma sentença incompleta, baseando-se no contexto e em palavras anteriores (OLIVEIRA et al., 2022). Para a tarefa de Sumarização, recebe-se um texto longo como entrada com o objetivo de resumir e gerar uma versão curta desse texto, mantendo os fatos mais importantes (TUNSTALL; WERRA; WOLF, 2022).

A Figura 4 apresenta um resumo da evolução do uso do PLN e AP ao longo do tempo. No início dos anos 2000, o AP passou a ser usado em tarefas de modelagem de linguagem, cujas características são apresentadas na Seção 2.6. Segundo Caseli e Nunes (2023), os modelos de linguagem neurais relacionados a RNPs iniciaram com os estudos de Xu e Rudnicky (2000) e Bengio, Ducharme e Vincent (2000). A partir de 2013, foram desenvolvidas novas técnicas de AP utilizando *word embbedings*, como o algoritmo *Word2Vec*, proposto por Mikolov et al. (2013). *Word embbedings* são vetores numéricos densos, que representam semanticamente cada palavra em um espaço vetorial (SANTOS et al., 2022). Com o *Word2Vec*, um novo modelo de representação vetorial semântica foi apresentado, demonstrando que palavras semelhantes estariam próximas umas às outras em um espaço vetorial (OLIVEIRA et al., 2022). O *Word2Vec* possui duas arquiteturas, uma chamada de *Continuos Bag of Words* (CBOW) e a outra chamada de *Skip-Gram*. Na arquitetura CBOW, o modelo *Word2Vec* recebe como entrada uma janela de palavras para tentar prever qual a palavra apareceria para o determinado contexto. Para a arquiterura *Skip-Gram* a entrada é uma palavra, e o modelo *Word2Vec* tenta prever as palavras vizinhas (MIKOLOV et al., 2013).

Ainda em 2013, modelos de RNP também passaram a serem utilizados em tarefas de PLN como Redes Neurais Recorrentes (do inglês *Recurrent Neural Networks* (RNN)) e as Redes Neurais Convolucionais (do inglês *Convolutional Neural Networks* (CNN)). Em 2014, uma abordagem chamada de *sequence-to-sequence learning* (seq2seq) foi proposta por Sutskever, Vinyals e Le (2014). A ideia do *seq2seq* é utilizar uma RNP para fazer o mapeamento de uma sequência de tokens para outra como, por exemplo, na tarefa de tradução de textos (OLIVEIRA et al., 2022). No ano seguinte, Bahdanau, Cho e Bengio (2014) desenvolveu o Mecanismo de Atenção. A proposta teve como objetivo criar um método que utilizasse RNP para simular como os seres humanos tratam textos, colocando atenção em trechos específicos em uma sentença de cada vez e não no texto todo ao mesmo tempo (OLIVEIRA et al., 2022).

Figura 4 – Linha do tempo com avanços do PLN e AP.



Fonte: Adaptado de Oliveira et al. (2022).

Atualmente, os *Large Language Models (LLMs)* são as grandes inovações no campo do PLN. Os LLMs foram propostos em 2015 por Dai e Le (2015), mas apenas a partir de 2018, que esses modelos ganharam notoriedade para várias tarefas de PLN como a geração de textos, sumarização e classificação de textos (OLIVEIRA et al., 2022). Os LLM são descritos na Seção 2.6.4.

2.6 Modelos de Linguagem

Os Modelos de Linguagem podem ser probabilísticos ou neurais, sendo utilizados para construir representações numéricas de textos, de forma que essas representações possam capturar a semântica, contexto e relações entre as palavras, sendo utilizados tanto para gerar quanto para consumir textos mapeados para representações numéricas (CASELI; NUNES, 2023).

2.6.1 Modelos probabilísticos

Segundo Caseli e Nunes (2023), os modelos de linguagem probabilísticos atribuem uma probabilidade a uma sequência de palavras. Cada palavra ou subpalavra, chamada de *token*, compõe o vocabulário que o modelo conhece. Baseado nesse vocabulário, nas palavras anteriores e na distribuição de probabilidade que o modelo aprendeu, o próximo token de uma sequência é gerado. Quanto maior e mais diversificado o conjunto de textos maiores são as possibilidades dele possuir variações, isso também torna o processamento mais demorado.

Os modelos de linguagem probabilísticos são baseados em métodos de aprendizado estatístico (ZHAO et al., 2023). O conceito básico é criar um modelo para prever palavras de acordo com a suposição Markov e Schorr-Kon (1962), que sugere prever a próxima palavra baseando-se no contexto das palavras mais recente.

Conforme Zhao et al. (2023), modelos de linguagem probabilísticos podem ter comprimento de contexto fixo, sendo chamados de modelos de linguagem n-gram. Exemplo de modelos de linguagem n-gram são o bigrama e trigrama.

2.6.2 Modelos neurais

Os modelos de linguagem neurais utilizam as redes neurais como, por exemplo as RNNs e *Transfomers*, cuja arquiterura é detalhada na Seção 2.7, para construir representações numéricas de forma dinâmica, de acordo com o contexto em que as palavras estão inseridas (CASELI; NUNES, 2023). Ainda, segundo Caseli e Nunes (2023), nesse tipo de modelo, a rede neural tem seus pesos treinados para aprender a função de probabilidade do modelo de linguagem para prever a próxima palavra baseada nas palavras anteriores.

Utilizando os modelos neurais vários métodos passaram a produzir representações numéricas dinâmicas, baseadas no contexto que a sentença está inserida, sendo os chamados *embeddings* contextualizados (CASELI; NUNES, 2023). Para ilustrar os embeddings contextualizados, consideram-se as seguintes sentenças:

- 1. Ele sentia que a luz no fim do túnel da depressão estava distante;
- 2. A luz da janela não conseguia iluminar a escuridão da sua mente;
- 3. A luz do sorriso dos amigos parecia perdida.

Percebe-se que a palavra "luz" está associada a três significados diferentes. Na sentença 1, a palavras "luz" está relacionada a esperança, na segunda sentença o significado tem a ver com a claridade do sol, enquanto que, na sentença 3, a palavra representa o brilho ou alegria. Os modelos de linguagem neurais são capazes de gerar embeddings diferentes para cada um das sentenças. Dessa forma, permitem capturar também a semântica e contexto. Cabe salientar que os embeddings gerados serão diferentes para uma palavra, mesmo que ela possua o mesmo significado em outra sentença. Por exemplo, na sentença "Ele percebeu que a luz interior foi perdida por causa da depressão", a palavra "luz", nesse contexto, significa esperança, mesmo significado que aparece na sentença 1, ainda sim, baseado nas palavras anteriores, os modelos neurais produzem embeddings diferentes.

O uso de embeddings contextualizados em tarefas de PLN compreendem duas características: a geração de embeddings e a sua utilização em tarefas finais como, por exemplo, a classificação de textos (CASELI; NUNES, 2023). Alguns métodos para geração de embeddings são as RNNs como ELMo (*Embeddings from Language Models*) (PETERS et al., 2018), os *Transformers* como o GPT (*Generative Pre-trained Transformer*) (BROWN et al., 2020) e o BERT(DEVLIN et al., 2019), cuja explicação é descrita na Seção 2.8.

2.6.3 Modelos pré-treinados

Modelos de Linguagem pré-treinados dizem respeito à técnica de treinar RNP com uma quantidade significativa de textos sem nenhuma anotação, com o propósito de produzir um modelo que consiga compreender uma linguagem ou um texto (CASELI; NUNES, 2023). O pré-treinamento dos modelos de linguagem demonstrou melhorar tarefas de PLN como, por exemplo, a tarefa de reconhecimento de entidades nomeadas, perguntas e respostas e inferência de linguagem natural (DEVLIN et al., 2019). De acordo com Caseli e Nunes (2023), de forma geral o pré-treinamento de um modelo de linguagem segue as seguintes etapas:

- Seleção do *corpus*: escolher o *corpus* adequado de acordo com o domínio que se deseja gerar o modelo;
- Pré-processamento do texto: apesar dos modelos de linguagem não precisarem de muitas atividades de pré-processamento, é recomendável que o *corpus* selecionado passe por atividades como remoção de URLs, caracteres especiais e códigos HTML;
- Tokenização: nesta etapa, o intuito é dividir o texto em unidades de palavras e subpalavras, podendo utilizar um tokenizador já treinado ou realizar o pré-treino do tokenizador;
- Determinar a arquitetura do modelo: escolher qual modelo usar para gerar o modelo pré-treinado. Na presente pesquisa, o BERT foi o modelo definido.
- Definir a tarefa alvo: selecionar a tarefa com a qual o modelo de linguagem será pré-treinado como, por exemplo, a tarefa de *Masked LM* e *Next Sentece Prediction* utilizada no pré-treino do BERT (ver Seção 2.8);
- Definição de hiperparâmetros: determinar as configurações de parâmetros como taxa de aprendizado, número de épocas, tamanho da amostra do lote de treinamento;
- Avaliação: avaliar o modelo em tarefas de PLN como a tarefa de classificação de textos.

Particularmente, com relação à tokenização, um vocabulário de palavras e subpalavras é produzido com características sobre o contexto e idioma a partir do *corpus* utilizado (TUNSTALL; WERRA; WOLF, 2022). O WordPiece(WU et al., 2016) e o SentencePiece (KUDO; RICHARDSON, 2018) são exemplos de tokenizadores. Na tokenização com WordPiece, o texto é inicialmente dividido em palavras, em seguida, essas palavras podem ser segmentadas em unidades de subpalavras. Essa divisão permite que o tokenizador possa realizar uma combinação de subpalavras evitando palavras desconhecidas no vocabulário criado. As palavras segmentadas recebem o prefixo "##"indicando a continuação da palavra. A tokenização do SentencePiece divide o texto substituíndo os caracteres de espaço em branco pelo meta-símbolo "▁"(U+2581) (SOUZA; NOGUEIRA; LOTUFO, 2020). Por exemplo, supondo que para a

sentença "Fui diagnosticado com depressão pelo psiquiatra e estou muito preocupado", as palavras "preocupado"não esteja no vocabulário, a tokenização utilizando o Wordpiece poderia ser: 'Fui' 'diagnosticado' 'com' 'depressão' 'pelo' 'psiquiatra' 'e' 'estou' 'muito' 'pre' '##ocupado'. Já utilizando o SentencePiece a sentença indicada ficaria:'▁Fui' '▁diagnosticado' '▁com' '▁depressão' '▁pelo ' '▁psiquiatra' '▁e' '▁estou' '▁muito' '▁ pre' 'ocupado'. Para compreender a atuação do tokenizador na prática, ainda considerando a sentença *"Fui diagnosticado com depressão pelo psiquiatra e estou muito preocupado"*. Após a tokenização, como ilustrado na Tabela 1, a sentença poderia ser traduzida para uma representação de IDs numéricos da seguinte forma: [101, 1704, 16008, 2512, 2587, 1312, 2904, 13004, 511, 2403, 84, 102, 0, 0, 0, 0, 0, 0 , 0,....0]. Onde os IDs 101 e 102 indicam os tokens de controle [CLS] e [SEP], respectivamente, e os outros IDs representam as palavras e subpalavras da sentença. Nota-se que a sequência é completada com zeros (0), token de controle [PAD], para que a sua representação complete o tamanho de 512 tokens. Considerando que a palavra "preocupado"não exista no vocabulário, mas a palavra "pre"e "ocupado"existam, o tokenizador faz a divisão evitando atribuir o token de controle [UNK] e permite a combinação de subpalavras conhecidas. A palavra "preocupado"foi dividida nas subpalavras "pre"e "##ocupado", cada uma com um ID diferente. Os caracteres "##"indicam a continuidade da palavra anterior. O ID 101 aponta o início da sentença, já o ID 102, neste caso, indica o fim da sentença.

Tabela 1 – Representação da sentença e seus respectivos IDs.

[CSL]	Fui	diagnosticado	com	depressão	pelo	psiquiatra	e	estou	pre	##ocupado	[SEP]	[PAD]
101	1704	16008	2512	2587	1312	2904	13004	511	2403	84	102	0

Fonte: Elaborado pelo autor.

Após toda as etapas do pré-treino ser concluído, os modelos de linguagem podem ser ajustados através do *Fine tuning* (ajuste fino), transferindo o conhecimento aprendido durante o pré-treino para tarefas de PLN como classificação de texto, sumarização, geração de texto (OLIVEIRA et al., 2022). Os modelos de linguagem pré-treinados como o BERT produzem representações numéricas das palavras de acordo com o contexto as quais elas estão inseridas, permitindo que esses recursos possam ser aproveitados para uma tarefa final como, por exemplo, classificar se uma postagem do Reddit possui indícios de depressão.

Como dito anteriormente, uma das formas de avaliar o desempenho de um modelo pré-treinado é utilizá-lo em uma tarefa final de PLN, como a classificação de texto. Nesse cenário, as medidas precisão, revocação e *F1-score* são calculadas de acordo com as Equações 1, 2 e 3, respectivamente. A precisão representa a proporção de classificações corretas para uma classe, já a revocação avalia se existem muitos falsos negativos em comparação com a quantidade de verdadeiros positivos. Já a *F1-score* é a média harmônica da precisão e revocação.

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

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

$$\text{F1} = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3}$$

Modelos de Linguagem podem ser baseados em *Transformer* cuja arquitetura é explicada na Seção 2.7.

2.6.4 LLM

Os LLMs são modelos de linguagem treinados com um enorme quantidade de dados não rotulados com o objetivo de aprender sobre contextos, idioma, ou ainda, aprender sobre um domínio de dados específico como, por exemplo, o domínio da depressão (CASELI; NUNES, 2023). Em relação aos modelos de linguagem pré-treinados, os LLMs diferenciam-se pois possuem, normalmente, bilhões ou até centenas de bilhões de parâmetros (ZHAO et al., 2023). Outra diferença é que os LLMs têm como principal função a geração de textos, fazendo parte da categoria da IA generativa (CASELI; NUNES, 2023).

Os LLMs podem ser utilizados a partir de instruções definidas através de linguagem natural utilizando os *prompts* (CASELI; NUNES, 2023). Esta habilidade, chamada de aprendizado em contexto (em inglês, *in-context learning* ou *few-shot prompt*), de acordo com Brown et al. (2020), é demonstrada quando esses modelos realizam tarefas para as quais não foram treinados especificamente.

Alguns exemplos de LLMs são o GPT-3 e GPT-4 (BROWN et al., 2020), utilizados para conversação através do ChatGPT[2] o LLaMA (TOUVRON et al., 2023), o PaLM (CHOWDHERY et al., 2023) e o Sabiá (PIRES et al., 2023), este na língua portuguesa e utilizado para conversação pelo Maritalk[3].

2.7 *Transformer*

Durante a evolução do PLN e AP, como foi mostrada na Figura 4, surgiram os *Transformers*. *Transformer* é uma arquitetura baseada em RNP que utiliza uma estrutura codificador-decodificador em conjunto com um mecanismo de atenção, permitindo realizar tarefas de PLN como, por exemplo, tradução de textos (VASWANI et al., 2017). O codificador é formado principalmente por um mecanismo com várias cabeças de autoatenção (*multi-head self-attention*) e uma rede feed-forward. Segundo Tunstall, Werra e Wolf (2022), uma rede feed-forward é uma rede neural simples com duas camadas totalmente conectadas que processam cada incorporação (*Embedding*) de forma independente. A estrutura do decodificador é semelhante ao codificador,

[2] https://chat.openai.com/
[3] https://chat.maritaca.ai/

entretanto, no decodificador é adicionada uma terceira subcamada com o objetivo de aplicar o mecanismo de atenção sobre a saída do codificador, conforme pode ser observado na Figura 5 (VASWANI et al., 2017). A subcamada com o mecanismo de autoatenção possibilita atribuir um peso para cada elemento de um vetor com a representação numérica do texto.

De acordo com (LIN et al., 2022), a arquitetura Transformer pode ser utilizada de três formas:

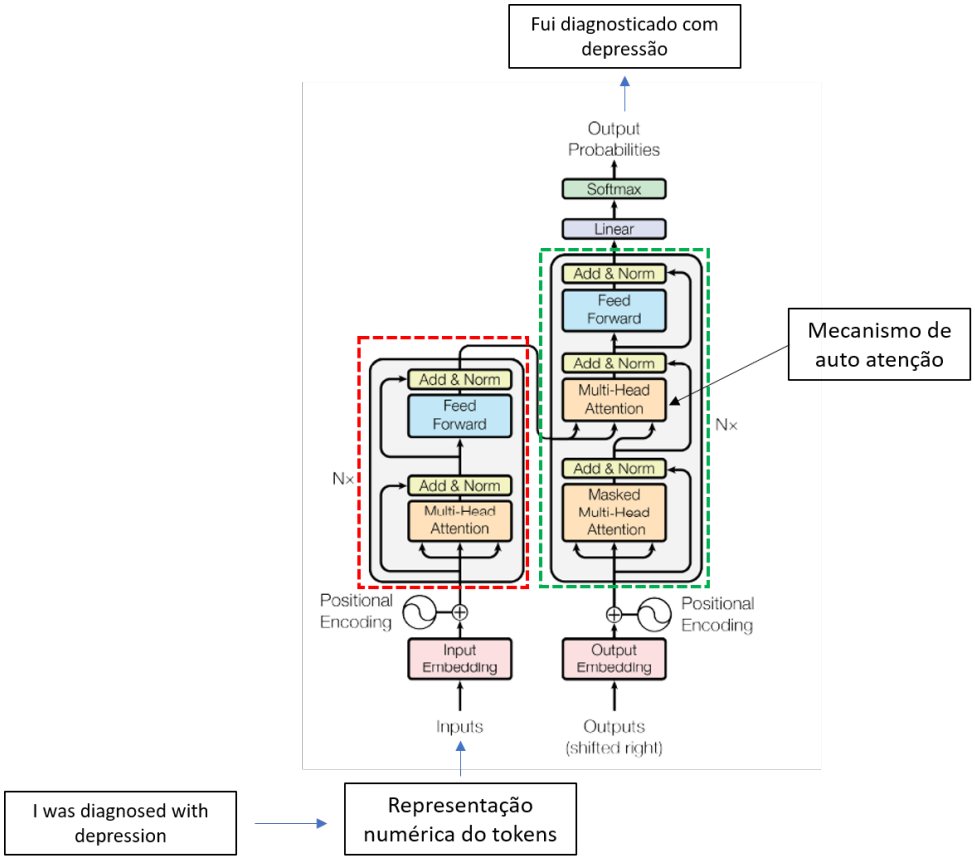
- Codificador-Decodificador: Utiliza a arquitetura completa do Transformer e é usada, geralmente, para tarefas como tradução de texto;
- Apenas codificador: As saídas do codificador são utilizadas como representação da sequência de entrada, aplicadas em tarefas de classificação de textos;
- Apenas decodificador: Só é utilizado o decodificador da arquitetura Transformer, geralmente em tarefas como, por exemplo, geração de texto.

A Figura 5 mostra a arquitetura geral de um *transformer*. No retângulo pontilhado em vermelho da figura, o codificador recebe como entrada uma representação numérica da sentença "I was diagnosed with depression", para gerar representações contextuais para cada token de entrada. O decodificador, retângulo pontilhado em verde, recebe as representações contextuais geradas pelo codificador para produzir os tokens de saídas em tarefas como, por exemplo, geração de textos ou tradução automática de textos. Com o surgimento dos *transformers* muitos modelos de linguagem pré-treinados surgiram, como, por exemplo, o BERT (ver Seção 2.8).

2.8 BERT

BERT é um acrônimo para *Bidirectional Encoder Representations from Transformers*, um modelo de linguagem desenvolvido por Devlin et al. (2019) baseado na arquitetura *Transformer*. Diferentemente de outros modelos de linguagem como OpenAI GPT (RADFORD et al., 2018), que aprende o contexto e gera representações das palavras considerando apenas uma direção (da esquerda para direita), o BERT foi desenvolvido para aprender e gerar representações de forma bidirecional (da esquerda para a direita e da direta para a esquerda). Isso permite compreender melhor o contexto de uma palavra já que o modelo é capaz de gerar representações analisando o termo anterior e posterior dessa palavra em uma sentença. O BERT foi implementado em duas versões de tamanho: *Base* e *Large*. A versão *Base* foi configurada com 12 camadas de *transformers*, 12 cabeças de atenção (*self attention*) e *embeddings* com dimensão igual a 768, totalizando 110 milhões de parâmetros. Já a versão *Large* foi configurada com 24 camadas de *transformers*, 16 cabeças de atenção e *embeddings* com dimensão igual a 1024, totalizando 340 milhões de parâmetros. Para realizar a etapa de pré-treino do BERT foram utilizados o *corpus* do Wikipédia no idioma inglês com 2.500 milhões de palavras e o BooksCorpus que possui 800

Figura 5 – Arquitetura do *Transfomer*.



Fonte: Adaptado deVaswani et al. (2017).

milhões de palavras (ZHU et al., 2015). O modelo utilizou um vocabulário de 30 mil *tokens* que foi gerado a partir do tokenizador WordPiece (WU et al., 2016).

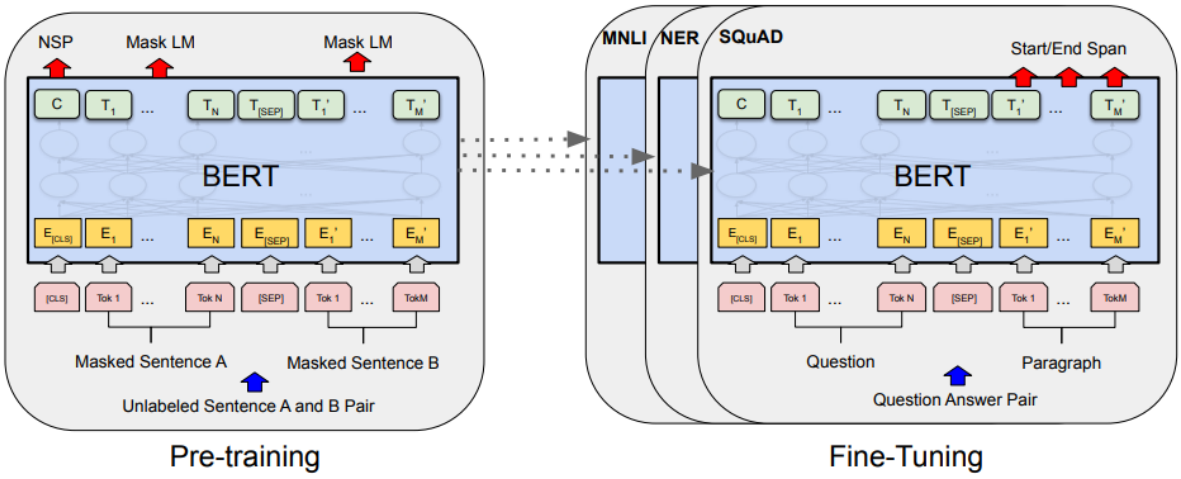
O BERT foi implementado em duas etapas: Pré-treino e Ajuste fino (*Fine tuning*). Na etapa do pré-treino o modelo é treinado com dados não rotulados para aprender sobre o contexto utilizando duas tarefas não supervisionadas: *Masked Language Modeling (Masked LM)* e *Next Sentence Prediction* (NSP). A tarefa de *Masked LM* consiste em mascarar de forma aleatória um percentual dos tokens de entrada e depois realizar a previsão dos tokens mascarados. Os tokens são substituídos pelo token "[MASK]". No pré-treino do BERT foram mascarados aleatoriamente 15% dos tokens de entrada. A tarefa de NSP consiste em pré-treinar o modelo BERT para que ele seja capaz de predizer, dadas as sentenças A e B, se a sentença B é continuação da sentença A .

Na segunda etapa do BERT, o ajuste fino, o modelo é iniciado com os valores dos parâmetros do pré-treino e depois ajustados utilizandos dados rotulados para tarefas como classificação de texto, reconhecimento de entidade nomeadas, ou perguntas e respostas.

A Figura 6 ilustra uma visão geral do pré-treino e ajuste fino do BERT. No lado esquerdo é mostrado a arquitetura do pré-treino utilizando dados não rotulados para o treinamento do modelo e as tarefas de *Masked LM* e NSP. O token "[CLS]"é adicionado no início de cada

sentença, e o token "[SEP]" indica a mudança entre duas sentenças ou o fim da sentença, sendo utilizados em ambas as tarefas. No lado direito, a Figura 6 mostra que o ajuste fino inicia o modelo com os parâmetros do pré-treino para realizar tarefas como perguntas e respostas.

Figura 6 – Pré-treinamento e Ajuste fino do BERT.



Fonte: Devlin et al. (2019).

2.9 Considerações

Este capítulo apresentou os principais conceitos que nortearam esta pesquisa. A seguir, o Capítulo 3 descreve os trabalhos relacionados à proposta desta dissertação, mostrando modelos de linguagem referentes à depressão, apresentando um comparativo entre eles e indicando alguns diferenciais em relação à proposta desenvolvida na presente pesquisa, abordada no Capítulo 4.

3 TRABALHOS RELACIONADOS

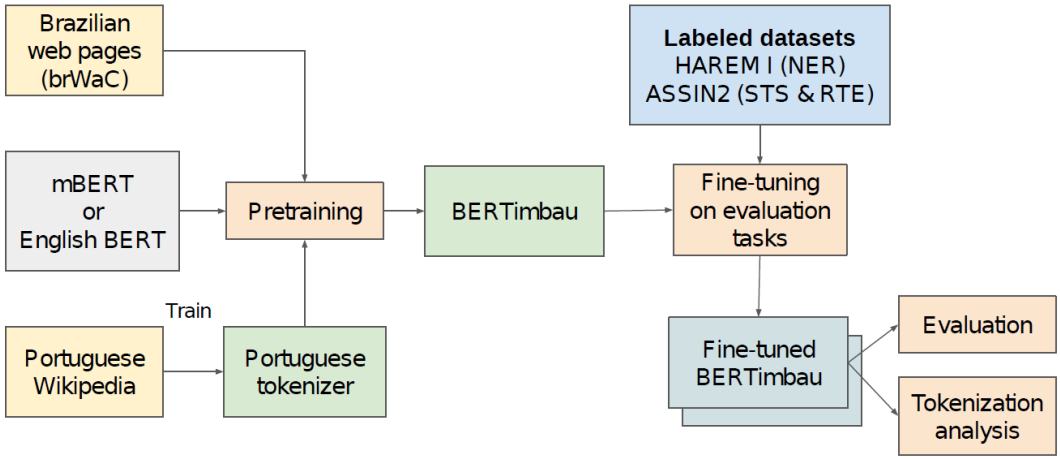
Neste capítulo são apresentados alguns trabalhos relacionados que utilizaram pré-treinamento de modelos de linguagem, seja para o domínio da depressão ou para um domínio mais geral no idioma inglês e português. Adicionalmente, é descrito um trabalho que construiu um conjunto de dados para depressão e ansiedade em português. Por fim, o capítulo analisa de modo comparativo os trabalhos descritos.

3.1 BERTimbau: pretrained BERT models for Brazilian Portuguese

Souza, Nogueira e Lotufo (2020) desenvolveram o BERTimbau, um modelo de linguagem pré-treinado a partir do BERT para o idioma português brasileiro com domínio geral. Para realizar o pré-treino do BERTimbau os pesquisadores geraram um vocabulário a partir da tokenização de sentenças da Wikipédia em português em domínio geral. O vocabulário foi gerado com 30.000 subpalavras. Em seguida, utilizaram como entrada de dados para o pré-treino o *corpus* brWaC (Brazilian Web as Corpus) (FILHO et al., 2018), um conjunto de dados que inclui conteúdo textual de páginas web brasileiras na internet que possui 145 milhões de sentenças. Após realizar o pré-treino, o BERTimbau foi avaliado em 3 tarefas de PLN, a saber: Similaridade Textual de Sentenças, Reconhecimento de Implicação Textual e Reconhecimento de Entidades Nomeadas. Essas tarefas foram escolhidas para que o BERTimbau fosse avaliada em tarefas em nível de sentença e tarefas em nível de token. Os pesquisadores destacam também que essas tarefas de PLN foram escolhidas por terem disponíveis conjuntos de dados rotulado para serem usados como *baseline*. A Figura 7 mostra uma abtração da arquitetura e das tarefas utilizadas na avaliação do BERTimbau. Os pesquisadores pré-treinaram o BERTimbau nas arquiteruras Base e *Large*. Para a arquitetura Base o pré-treinamento foi inicializado a partir do mBERT, um modelo multilingue do BERT com 104 idiomas. Já o pré-treinamento do BERTimbau *Large* foi inicializado a partir do BERT, isso porque o mBERT é disponibilizado apenas na versão Base. O brWaC *corpus* juntamente com o vocabulário são passados para o pré-treinamento do BERTimbau. Depois do pré-treino, o BERTimbau passa por ajuste para ser utilizado em tarefas de PLN mencionadas anteriormente.

A tarefa de similaridade textual de sentenças consiste em medir o quão uma sentença é semelhante a outra semanticamente. Sentenças totalmente diferentes recebem rótulos com valores mais próximos de 1. A tarefa de Reconhecimento de Implicação Textual é uma tarefa de classificação para predizer se, dada uma sentença A,a partir dela, pode-se deduzir uma sentença B, como pode ser observado na Figura 8 (SOUZA; NOGUEIRA; LOTUFO, 2020). Para as tarefas mencionadas, os pesquisadores do BERTimbau utilizaram o *corpus* ASSIN2 do trabalho de Real, Fonseca e Oliveira (2020), que possui 10.000 pares de sentenças rotuladas. A Figura

Figura 7 – Abstração da Arquitetura do BERTimbau.



Fonte: Souza, Nogueira e Lotufo (2020).

8 mostra exemplos de sentenças do conjunto de dados ASSIN2. Os rótulos para a tarefa de similaridade textual são valores contínuos que variam em uma escala de 1 a 5, de forma que, quanto mais similares semanticamente os pares de sentenças, mais esses são próximas de 5, já os rótulos da tarefa de Reconhecimento de Implicação Textual são "Entailment"(indicando que a sentença B é consequência da sentença A) ou "None"(quando a sentença B não pode ser deduzida a partir da sentença A).

A tarefa de Reconhecimento de Entidades Nomeadas consiste em identificar no texto entidades e classificá-las em categorias previamente definidas como localização, pessoa, organização (SOUZA; NOGUEIRA; LOTUFO, 2020). Os autores utilizaram o conjunto de dados *First HAREM* para treino e *MiniHAREM* para teste (SANTOS et al., 2006). Esses conjunto de dados tinham documentos de domínios de conhecimento variados e rótulos para 10 classes de entidades nomeadas, a saber: Pessoa, Organização, Local, Valor, Data, Título, Coisa, Evento, Abstração e Outros. Foram averiguados dois cenários diferentes, o primeiro levou em consideração todas as 10 classes, já no segundo cenário, apenas 5 classes (Pessoa, Organização, Local, Valor e Data) foram consideradas.

Para comparar e analisar se um modelo monolíngue, como o BERTimbau, teria desempenho melhor em tarefas como Similaridade textual de Sentenças e Reconhecimento de implicação textual, em relação a um modelo multilingue, os pesquisadores treinaram o mBERT. O mBERT foi treinado originalmente com artigos da Wikipédia em 104 idiomas. Os resultados mostraram que um modelo pré-treinado para um único idioma, ou seja, monolíngue, consegue obter melhores resultados nas tarefas avaliadas quando comparados a modelos multilíngues. Uma possível vantagem é que o modelo monolíngue como o BERTimbau possui um vocabulário que varia entre 30.000 a 50.000 *tokens*, já o vocabulário de um modelo multilíngue como o mBERT tem um vocabulário de 120.000 *tokens* para englobar 104 idiomas. Dessa forma,

Figura 8 – Amostras do conjunto de dados ASSIN2.

Gold STS/RTE	Sentence pair
5.0 / Entailment	A: *Os meninos estão de pé na frente do carro, que está queimando.* B: *Os meninos estão de pé na frente do carro em chamas.*
	English translation: A: The boys are standing in front of the car, which is burning. B: The boys are standing in front of the burning car.
4.0 / Entailment	A: *O campo verde para corrida de cavalos está completamente cheio de jóqueis.* B: *Os jóqueis estão correndo a cavalos no campo, que é completamente verde.*
	English translation: A: The green field for horse races is completely full of Jockeys. B: The Jockeys are racing horses on the field, which is completely green.
3.0 / Entailment	A: *A gruta com interior rosa está sendo escalada por quatro crianças do Oriente Médio, três meninas e um menino.* B: *Um grupo de crianças está brincando em uma estrutura colorida.*
	English translation: A: Four middle eastern children, three girls and one boy, are climbing on the grotto with a pink interior. B: A group of kids is playing in a colorful structure.
2.0 / None	A: *Não tem nenhuma pessoa descascando uma batata.* B: *Uma pessoa está fritando alguma comida.*
	English translation: A: There is no one peeling a potato. B: A person is frying some food.
1.0 / None	A: *Um cachorro está correndo no chão.* B: *A menina está batucando suas unhas.*
	English translation: A: A dog is running on the ground. B: The girl is tapping her fingernails.

Fonte: Souza, Nogueira e Lotufo (2020).

quando se utiliza o mBERT para um idioma específico, o tamanho do vocabulário será, na maioria das vezes, menor que o vocabulário de um modelo monolíngue.

3.2 MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare

No trabalho de Ji et al. (2022) foram desenvolvidos o MentalBERT e MentalRoBERTa, dois modelos de linguagens pré-treinados para o domínio da saúde mental no idioma inglês. Os dados para o pré-treinamento do MentalBERT e MentalRoBERTa foram obtidos de subcomunidades do Reddit relacionadas à saúde mental, são elas: *r/depression, r/SuicideWatch, r/Anxiety, r/offmychest, r/bipolar, r/mentalillness/,r/mentalhealth*. Com isso, os pesquisadores criaram um *corpus* com mais de 13 milhões de sentenças.

Após o pre-treinamento os autores do MentalBERT e o MentalRoBERTa os pesquisadores realizam o ajuste fino (*fine tuning*) dos modelos desenvolvidos para a tarefa de classificação de textos. Neste caso, detecção de transtornos mentais como a depressão, transtorno de ansie-

dade, ideação suicida e estresse. Para cada transtorno mencionado foram utilizados diferentes *corpus*, específicos para o domínio do transtorno mental. Em especial, para a depressão, foram usados os conjunto de dados dos trabalho de Coppersmith et al. (2015) com perfis de usuário com depressão no Twitter, Losada e Crestani (2016) com dados eRisk (*Early Risk Prediction on the Internet*), uma competição para detectar de forma precoce riscos à saúde mental, por fim, o trabalho de Pirina e Çöltekin (2018) com dados do Reddit sobre depressão.

Para avaliar os modelos desenvolvidos os autores do MentalBERT e MentalRoBERTa treinaram para a tarefa de classificação os modelos BERT e RoBERTa, pre-treinados originalmente com dados de domínio geral, e modelos de domínio específicos de saúde, neste caso o BioBERT(LEE et al., 2020), modelo pré-treinado para o domínio biomédico, e o ClinicalBERT (ALSENTZER et al., 2019), modelo pré-treinado com anotações clínicas. Todos esses modelos foram treinado com os *corpus* utilizados para o ajuste fino do MentalBERT e MentalRoBERTa.

Os resultados mostraram que para a tarefa de detecção de depressão,os modelos pré-treinamento com os dados nos domínios biomédico e clínico não foram tão relevantes. Em comparação com o BERT e o RoBERTa, o MentalBERT e MentalRoBERTa superaram, de acordo com as métricas de avaliação revocação e f1 os modelos com dados de domínio geral. Dessa forma, o pré-treinamento de modelos de linguagem para um domínio específico pode ser bastante competitivo para detectar sinais de depressão.

3.3 Detecting signs of depression from social media text using RoBERTa pre-trained language models

Liu et al. (2019) desenvolveram o modelo pré-treinado denominado RoBERTa, o qual consiste em replicar o treinamento do BERT com hiperparâmetros modificados e apenas com a tarefa de mascaramento de tokens (Masked LM) durante o treinamento. Com base no modelo RoBERTa, Poświata e Perełkiewicz (2022) criaram o DepRoBERTa, um modelo de linguagem pré-treinado para o domínio da depressão utilizando o *corpus* Reddit Mental Health Dataset (LOW et al., 2020), juntamente com um *corpus* obtido do Kaggle[1] com postagens no idioma inglês a partir de subcomunidades (subreddits) sobre depressão e suicídio no Reddit.

Inicialmente, os pesquisadores do DepRoBERTa utilizaram modelos de linguagem pré-treinados com dados de domínio geral como BERT, RoBERTa e XLNet. Entretanto, por serem de domínio mais abrangente, o *corpus* utilizado por esses modelos podem conter poucas informações sobre sintomas de depressão. Dessa forma, os autores do DepRoBERTa realizaram um pré-treinamento a partir do modelo RoBERTa, porém utilizando dados específicos do domínio da depressão.

O DepRoBERTa foi então usado como modelo na competição *Shared Task on Detecting Signs of Depression from Social Media Text* (KAYALVIZHI et al., 2022), cujo desafio era

[1] https://www.kaggle.com/xavrig/reddit-datasetrdepression-and-rsuicidewatch

criar um modelo para classificar postagens do Reddit, no idioma inglês, em uma das seguintes categorias:

- Sem depressão - postagens que apresentassem temas de propósito geral ou refletissem sentimentos momentâneos;
- Depressão moderada - postagens que refletissem mudança nos sentimentos(sentindos-se deprimido por algum tempo), poderações sobre ter esperança na vida;
- Depressão grave - postagens que demonstrassem mais de uma situação de transtorno ou postagens que indicassem histórico de tentativa de suicídio.

As postagens do dataset da competição foram coletadas de subcomunidades do Reddit tais como r/depressão, r/solidão, r/estresse e r/ansiedade. A rotulação das postagens foi realizada por dois especialistas do domínio que analisaram os textos e atribuíram os rótulos apropriados (SAMPATH; DURAIRAJ, 2022).

Os dataset da competição foi dividido pela organização em três conjuntos: treinamento, avaliação e testes. Os dados de treinamento e avaliação rotulados foram disponibilizados para que os participantes da competição desenvolvessem seus modelos. O conjunto de teste não possuia rótulos durante a competição, justamente para que os modelos desenvolvidos fossem avaliados em termos de acertos e erros na predição. Posteriormente os rótulos foram divulgados. A figura 9 ilustra uma amostra dos dados do dataset utilizado na competição. Basicamente, o dataset era constituído por três atributos: um identificador (PID) que sinaliza em qual partição a instância se encontra (treino, teste ou desenvolvimento) se os dados eram de treinamento, desenvolvimento ou teste; o texto da postagem (Text) e o rótulo atribuído a cada postagem (Label).

Figura 9 – Amostra do conjunto de dados.

PID	Text	Label
train_pid_6035	Happy New Years Everyone : We made it another year	not depression
train_pid_35	My life gets worse every year : That's what it feels like anyway....	moderate
train_pid_8066	Words can't describe how bad I feel right now : I just want to fall asleep forever.	severe

Fonte: Poświata e Perełkiewicz (2022).

Para ser utilizado na competição mencionada anteriormente, o DepRoBERTa foi ajustado *(Fine tuning)* para a tarefa de classificação de texto utilizando o conjunto de dados de treinamento e avaliação disponibilizados pela competição. Os resultados mostraram que um modelo pré-treinado para um domínio específico consegue extrair as características e detalhes da linguagem, neste caso no idioma inglês, melhor quando comparado a um modelo que teve em seu pré-treino dados de domínio geral. Isso porque, o DepRoBERTa obteve as melhores

avaliações em termos de precisão, revocação e F1. Portanto, treinar um modelo utilizando um conjunto de dados sobre depressão pode ajudar a classificar e identificar sintomas de depressão a partir de postagens em redes sociais de forma mais acertiva.

3.4 SetembroBR: a social media corpus for depression and anxiety disorder prediction

O SetembroBR foi um trabalho desenvolvido por Santos, Oliveira e Paraboni (2023). Neste trabalho, foram construídos dois conjuntos de dados em português, um para o domínio da depressão e outro para o de transtorno de ansiedade. Apesar de serem independentes, algumas instâncias podem estar em ambos os conjuntos, considerando o caso de um determinado usuário relatar mais de um dos transtornos. O objetivo da criação desse *corpus* é servir para estudo e desenvolvimento de modelos preditivos para detecção de depressão e do transtorno de ansiedade de perfis de usuário nas redes sociais.

Os dados foram coletados a partir de postagens no Twitter (atualmente chamado de X[2]) no período compreendido entre setembro de 2019 a fevereiro de 2021, utilizando strings de busca tais como "Comecei a tomar medicamentos antidepressivos" para identificar perfis classificados como "usuários diagnosticados". Usuários classificados como "usuários de controle"são perfis de usuários que atendem aos seguintes requisitos: (a) não apresentaram indícios de sintomas depressivos e/ou de transtorno de ansiedade; (b) possuem uma atividade de mais de 1.000 tweets postados; (c) não postou mensagem compatível com as strings de busca, e (d) não apresenta mais de 10.000 seguidores no Twitter. Perfis de usuários com mais de 10.000 seguidores foram excluídos para evitar contas de uso profissional. Também foram descartados usuários que relataram ter outros transtornos mentais, como transtorno bipolar, borderline, esquizofrenia, ou autismo. Os pesquisadores buscaram identificar a última postagem antes do diagnóstico relatado na linha do tempo nas postagens de cada usuário, pois esta foi a parte considerada útil para a previsão de depressão ou do transtorno da ansiedade. Algumas de string de busca enviadas à API(*Application Programming Interface*) do Twitter para coletar postagens relacionadas a sintomas de depressão ou sintomas do transtorno da ansiedade foram:

- comecei a tomar medicamento antidepressivo;
- acabei de ser diagnosticado depressão;
- eu fui diagnosticado ansiedade;
- antidepressivo prescrito;
- psicólogo diagnosticou ansiedade.

[2] https://twitter.com/

Para qualificar o uso dos *corpus* criados, os autores realizaram experimentos para predizer se um usuário estaria ou não com depressão, e se estaria com um transtorno de ansiedade ou não, de acordo com suas as postagens. Os experimentos também serviram para criar um *baseline* para estudos posteriores sobre a depressão e o transtorno de ansiedade no idioma português brasileiro. Para realizar os experimentos, os pesquisadores realizam uma classificação binária de perfis de usuários, classificando se um o perfil apresenta depressão ou não, bem como se um perfil de usuário apresenta transtorno de ansiedade ou não. Os autores do SetembroBR utilizaram como classificadores uma regressão logística, uma rede neural LSTM(*Long short term memory*), uma rede neural CNN(*Convolutional Neural Networks*) e o modelo de linguagem BERT. Para avaliação, os pesquisadores analisaram os resultados das métrica precisão, revocação e f1. Na comparação dos modelos experimentados, o BERT foi o que obteve os melhores resultados na predição de depressão e ansiedade.

3.5 BERTabaporu: Assessing a Genre-specific Language Model for Portuguese NLP

A pesquisa de Costa et al. (2023) desenvolveu o BERTabaporu, um modelo de linguagem BERT pré-treinado com dados no idioma português brasileiro. Para o pré-treino do modelo, os autores utilizaram um *corpus* de dados do Twitter contendo texto de tópicos relacionados a política, saúde mental e Covid-19. Na etapa de pré-processamento, foram removidos emoticons, caracteres não alfabéticos, textos em outro idioma e nomes de usuários.

Para fins de avaliação, o modelo BERTabaporu foi comparado com o modelo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) em três problemas de classificação binária, a saber:

- Previsão de postura: problema interessado em inferir se uma sentença demonstra ser favorável ou contra a um determinado tópico-alvo. Por exemplo, na sentença "Uma renda básica universal aliviaria a pobreza", a postura é favorável à "renda básica universal";
- Previsão de alinhamento político: visa inferir se uma pessoa apoia ou não uma orientação política (direita ou esquerda) no Brasil, com base em suas publicações em uma rede social.
- Previsão de estado de saúde mental: tem como objetivo determinar a partir de postagens de uma rede social se uma pessoa tem tendência a algum transtorno mental, como a depressão ou transtorno de ansiedade;

Particularmente, para a tarefa de previsão do estado de saúde mental, o autores do BERTabaporu realizaram a classificação em nível usuário, conforme a linha do tempo das postagens desses usuários. O *corpus* utilizado para esta tarefa foi o SetembroBR (SANTOS; OLIVEIRA;

PARABONI, 2023), descrito na Seção 3.4. Assim, o BERTabaporu realizou classificação para dois transtornos mentais, a saber: a depressão e o transtorno da ansiedade.

Após a realização dos experimentos em cada tarefa citada previamente, os resultados mostraram que o BERTabaporu conseguiu superar o BERTimbau em todas as tarefas, em termos de avaliação realizadas com as métricas precisão, revocação e f1. Em especial, para a tarefa de saúde mental, sugere que o conhecimento adquirido no pré-treinamento com dados de saúde mental permitiu ao BERTabaporu superar o BERTimbau para classificar perfis de usuário no Twitter diagnosticados com depressão e perfis de usuários diagnosticados com transtorno de ansiedade.

3.6 Síntese sobre os trabalhos relacionados

A Tabela 2 apresenta um resumo comparativo dos trabalhos discutidos. Para o domínio específico da depressão utilizando modelos pré-treinados, há apenas trabalhos no idioma inglês, como mostrados nas pesquisas de Ji et al. (2022) e Poświata e Perełkiewicz (2022). Reafirmando o que relatado na Revisão Sistemática da Literatura de Herculano et al. (2022). Apesar da pesquisa de Costa et al. (2023) ter pré-treinado um modelo no idioma português com dados de domínio geral englobando também dados de saúde mental, ele não é exclusivo do domínio da depressão. Da mesma forma, o trabalho de Souza, Nogueira e Lotufo (2020) desenvolveu um modelo pré-treinado em português brasileiro a partir de dados de domínio geral de páginas web.

O trabalho de Santos, Oliveira e Paraboni (2023) criou um *corpus* no idioma português brasileiro com postagens do Twitter (hoje denominado X) de pessoas que já haviam sido diagnosticada com depressão e/ou transtorno de ansiedade. Ou seja, os usuários relataram em seus perfis que estavam doentes. No entanto, detectar indícios de depressão a partir de postagens nas redes sociais antes de um diagnóstico clínico pode ajudar na busca por um tratamento com especialistas, antecipando a cura e até evitando que a doenças se agrave e possa levar à morte.

Assim, em comparação com os trabalhos relacionados citados anteriormente, nossa implementa um modelo pré-treinado baseado no modelo BERT, para o domínio específico da depressão usando postagens de redes sociais no idioma português brasileiro. Para isso, será construído um *corpus* a partir de postagens na língua portuguesa com teor depressivo no Reddit. Após o pré-treino, o modelo é adaptado para a tarefa de classificação de texto, classificando as postagens em três níveis, a saber: Sem depressão, Depressão moderada, Depressão grave.

Tabela 2 – Comparativo entre os trabalhos relacionados

Trabalho	**Domínio de dados**	**Realizou pré-treino**	**Fine Tuning**	**Idioma**	**Corpus**	**Contribuições**
MentalBert	Depressão	Sim	Sim	Inglês	Produziu	Desenvolveu dois modelos pré-treinados baseados no BERT e no RoBERTa para o domínio dos transtornos mentais no idioma inglês
DepRoBERTa	Depressão	Sim	Sim	Inglês	Utilizou existente	Desenvolveu um modelo pré-treinado em inglês baseado no RoBERTa, Classificou as postagens em nível de acordo com o nível de depressão
BERTimbau	Geral	Sim	Sim	Português	Utilizou existente	Desenvolveu um modelo pré-treinado em português baseado no BERT com o domínio de dados geral
SetembroBr	Depressão e Ansiedade	Não	Não	Português	Produziu	Construiu um corpus nos domínios da depressão e ansiedade em português
BERTabaporu	Geral com dados sobre saúde mental	Sim	Sim	Português	Produziu	Desenvolveu um modelo pré-treinado em português baseado no BERT com foco no gênero de textos das redes sociais

Fonte: Elaborada pelo Autor.

4 MODELO DE LINGUAGEM DEPREBERTBR

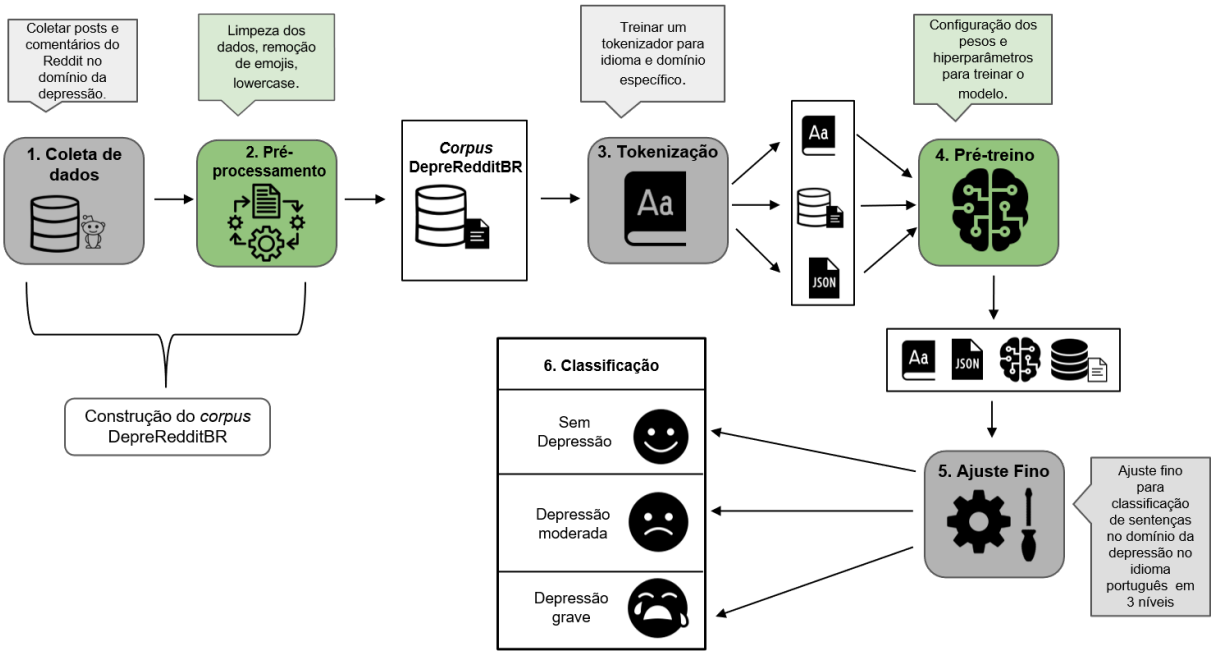
Este capítulo apresenta o modelo de linguagem DepreBERTBR e a metodologia empregada para a implementação da abordagem do modelo especializado com textos de postagens do Reddit com teor depressivo, no idioma português brasileiro.

O modelo DepreBERTBR proposto está inserido no contexto particular da depressão. Por ser um modelo especializado, este pode ser ajustado para utilização em tarefas específicas de PLN. Particularmente, o DepreBERTBR é treinado para um problema de classificação de postagens do Reddit em uma das seguintes classes: Sem depressão, Depressão moderada e Depressão grave. O DepreBERTBR é a base para resolução das questões de pesquisa relembradas a seguir:

- QP1: Como classificar postagens de redes sociais no idioma português do Brasil considerando um possível grau de depressão?
- QP2: Utilizar um modelo de linguagem pré-treinado no domínio específico da depressão no idioma português do Brasil pode ajudar a determinar o grau de depressão percebido em postagens de redes sociais?

A Figura 10 ilustra uma visão geral das etapas do processo de desenvolvimento do modelo DepreBERTBR. Cada etapa será abordada em maiores detalhes nas seções seguintes. Inicialmente, as etapas 1 (Coleta de dados) e 2 (pré-processamento de dados) combinadas resultam na construção de um *corpus* no idioma português brasileiro, com postagens específicas de subcomunidades no Reddit que abordam discussões sobre depressão. O *corpus* resultante é nomeado DepreRedditBR. A partir do DepreRedditBR, inicia-se o processo de tokenização na etapa 3, o qual resulta na determinação do vocabulário do modelo e um tokenizador, ambos no idioma português brasileiro e específico ao domínio da depressão. O vocabulário, o tokenizador e o *corpus* DepreRedditBR são utilizados como entrada para o pré-treinamento do DepreBERTBR (etapa 4). No pré-treino, o modelo aprende sobre o contexto determinado pelo DepreRedditBR. Com a conclusão da etapa 4, tem-se o produto DepreBERTBR que pode ser empregado para diversas finalidades (tarefas) como, por exemplo, classificação de textos. Na etapa 5 (ajuste fino), o DepreBERTBR é ajustado para o problema de classificação particular dentro do escopo desta dissertação, utilizando um *corpus* com postagens do Reddit previamente rotuladas por especialistas. O classificador, após ajustado, então, realiza a predição em uma das 3 (três) classes a seguir: Sem depressão, Depressão moderada e Depressão grave.

Figura 10 – Abordagem DepreBERTBR.



Fonte: Elaborado pelo autor.

4.1 Construção do *corpus* DepreRedditBR

Para a construção do *corpus*, nomeado DepreRedditBR, foi realizado um levantamento buscando identificar as redes sociais que oferecessem ambientes de discussão relacionadas ao tópico *depressão*. Ainda, a rede social deveria prover condições de extração de dados por meio de uma API própria. O Reddit demonstrou grande potencial nesse sentido, pois existem subcomunidades (subreddits) ativas em torno do domínio da depressão em que os usuários se sentem à vontade para expor seus sentimentos em postagens ou comentários.

Para realizar a coleta de dados, foi necessário desenvolver uma aplicação web denominada *Extrator* (ESTRELA et al., 2024), conveniente para extração de postagens do Reddit. A extração de postagens do Reddit considerou as seguintes definições:

- API de extração: a biblioteca PRAW[1] (*The Python Reddit API Wrapper*) permite acesso simplificado à API do Reddit;
- Subreddits coletados: r/arco_iris, r/desabafos, r/desabafo, r/relacionamentos, r/transbr, r/EuSouOBabaca, r/BissexualidadeBr, r/AnsiedadeDepressao, r/brasil, r/relatosdoreddit, r/brdev e r/PsicologiaBR. A lista dos subreddits foi determinada após examinação de várias subcomunidades candidatas, sendo aprovadas aquelas que majoritariamente contemplam relatos e ou comentários de caráter depressivo;

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

- String de busca: com base nos trabalhos de Azam et al. (2021) e Nascimento et al. (2018), foram examinados, avaliados e aprovados por especialistas de domínio, um conjunto de termos associados ao contexto da depressão, resultando na seguinte string de busca: *"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 melancolico OR melancolica OR melancolia OR cansado OR cansada OR sufocado OR sufocada"*

O Extrator teve como saída arquivos no formato CSV (Valores separados por vírgula, do inglês, *Comma-separated values*). Ressalta-se que o título de uma postagem pode ter até 300 caracteres, o que pode permitir inferir pelo título se a postagem apresenta uma temática depressiva. Com isso, o título também foi usado para construir o *corpus*. Ao final o Extrator produziu um *corpus* com 200.030 instâncias, sendo compostas por títulos, postagens e comentários. Considerando que o pré-treinamento de modelos de linguagem requer um *corpus* com um grande volume de dados textuais, adicionalmente foram incorporadas ao DepreRedditBR um total de 3.404 postagens relacionadas à depressão, extraídas também do Reddit, porém, disponibilizadas pela plataforma Kaggle[2]. As postagens incluem dados relativos ao título, corpo da postagem e comentários associados. Apesar da coleta de dados do Extrator e da integração com o conjunto de dados da plataforma Kaggle terem produzido um *corpus* inicial no idioma português brasileiro, com dados associados ao domínio da depressão, observou-se a necessidade de incrementar o tamanho do *corpus*. A Revisão Sistemática da Literatura realizada por Herculano et al. (2022) evidenciou a existência de muitos trabalhos de construção de modelos de linguagem que foram desenvolvidos utilizando conjuntos de dados no idioma inglês. Entretanto, devido à quantidade e ao tamanho dos textos das postagens serem consideráveis, torna-se inviável e muito custoso realizar a tradução manualmente. Por isso, foi utilizada a API do Google Translate para realizara a tradução do idioma inglês para o português brasileiro.

As postagens traduzidas passaram a integrar o *corpus* DepreRedditBr que foi tratado na etapa 2 (pré-processamento), que será detalhada na Seção 4.2.

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

Aumentar o tamanho do *corpus* justifica-se por dois fatores: (a) o treinamento de um modelo de linguagem exige uma grande massa de dados textual para aquisição de um bom vocabulário, entender melhor o significado das sentenças, identificar o contexto e realizar diferentes tarefas relacionadas à linguagem natural, e (b) existe uma limitação de conjuntos de dados focados no domínio da depressão no idioma português. Sendo assim, ampliar um *corpus* com mais dados é de grande importância para o pré-treinamento do modelo independentemente da tarefa alvo.

A ideia de agrupar conjuntos de dados distintos não é original. O trabalho de Poświata e Perełkiewicz (2022) reuniu dois *corpus* com postagens no Reddit com teor depressivo no idioma inglês para o desenvolvimento do seu modelo de linguagem. 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 para investigar como a pessoas com diferentes transtornos mentais foram afetadas durante a pandemia da COVID-19. O segundo *corpus* foi obtido da plataforma Kaggle[3], contendo postagens dos subreddits r/depression e r/SuicideWatch. Dessa forma, utilizando a API do Google Translate, foram traduzidas do idioma inglês para o idioma português brasileiro 338.139 postagens do Reddit.

4.2 Pré-processamento dos dados

Após a etapa de entrada de dados, deu-se início à tarefa de pré-processamento dos dados. Todos os dados do DepreRedditBR resultantes do serviço do Extrator e da tradução de conjuntos de dados foram utilizados como entrada nesta etapa. As ações de pré-processamento de dados contemplam a remoção de URLs (*Uniform Resource Locator*) contidas nos textos, emojis e emoticons. Apesar das limitações inerentes à quantidade de caracteres do título e comentários (300 caracteres e 10 mil caracteres, respectivamente) mencionados na Seção 2.1, o texto de uma postagem no Reddit é de tamanho ilimitado, sendo comum encontrar postagens com várias quebras de linha. Assim, visando melhorar a qualidade do texto e, evitar ruídos e inconsistência na etapa 3 (4.3), também foram removidas quebras de linhas. Outra atividade realizada na etapa de pré-processamento foi a remoção de postagens com marcações $'NaN'$[4], $'[removed]'$, e $'[deleted]'$. Particularmente, as inserções $'[removed]'$ e $'[deleted]'$ se referem a fragmentos de texto removidos pelos moderadores do subreddit. Por fim, foram removidas postagens em duplicidade utilizando a biblioteca Pandas [5].

Ao término da etapa de pré-processamento, o *corpus* resultante contempla um total de 509.675 mil postagens. A Tabela 3 mostra um extrato do *corpus* DepreRedditBR depois das tarefas de pré-processamento de dados. Percebe-se que o texto das postagens pode assumir tamanhos diferentes, com textos variados entre curtos e muito longos, e com conteúdo normalmente

[3] https://www.kaggle.com/datasets/xavrig/reddit-dataset-rdepression-and-rsuicidewatch
[4] Indicador de que em tal ponto havia um vídeo ou foto no texto.
[5] https://pandas.pydata.org/

Tabela 3 – 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.

Fonte: Elaborado pelo autor.

com teor depressivo.

4.3 Tokenização e vocabulário

Para o pré-treinamento de um modelo de linguagem é necessário que o *corpus*, depois de pré-processado, passe pelo procedimento de tokenização (Etapa 3 da Figura 10). Na tokenização o texto é dividido em partes menores chamadas de *tokens*. Cada token pode ser uma palavra ou subpalavra, associado a um identificador (ID) único no vocabulário criado. Todo o *corpus* é tokenizado e transformado em um vocabulário, imprescindíveis para entrada na etapa de pré-treinamento do modelo (etapa 4 da Figura 10).

Ressalta-se que, no desenvolvimento do DepreBERTBR, o *corpus* resultante das etapas 1 e 2 foi utilizado primeiramente para pré-treinar um tokenizador com textos do domínio da depressão no idioma português brasileiro. Entretanto, nem sempre é necessário treinar um tokenizador quando se dispõe de um tokenizador já treinado com um vocabulário próprio ou adequado. Na concepção do DepreBERTBR, considera-se um *corpus* especializado, no idioma português brasileiro, sendo, por isso, necessário produzir o próprio tokenizador para aprender a relação entre as palavras do domínio particular da depressão.

O tokenizador pré-treinado tomou como base o tokenizador WordPiece (WU et al., 2016), o mesmo tokenizador utilizado no treinamento do BERT. A configuração do pré-treinamento do tokenizador adotou as seguintes definições:

- Os parâmetros da quantidade máximo de palavras e subpalavras que o vocabulário foi definido em 99.999. Se for definido um vocabulário com uma quantidade pequena pode resultar em muitos tokens desconhecidos na tokenização dos textos, o que causaria a perda de contexto;
- A quantidade máxima de tokens para representação numérica de uma sentença foi de 512 tokens. A representação numérica é um vetor com os ids de cada palavra ou subpalavra da sentença, juntamento com os tokens de controle. Para modelos como o BERT, é necessário que as sequências das representações numéricas tenham vetor com mesmo

tamanho. O limite máximo de tokens em uma representação numérica de entrada que o BERT permite é 512 tokens. Se um sentença tokenizada ultrapassar essa quantidade, os tokens sobressalentes serão truncados e sequências menores serão completadas com IDs igual a zero (0).

No pré-treinamento do tokenizador também é gerado um arquivo de configuração que contém os *special_tokens*, isto é, os tokens de controle utilizados pelo BERT durante o treinamento que também possuem IDs, a saber:

- (ID=101) - [CLS]: indicativo de início da sentença de entrada;
- (ID=100) - [UNK]: representação de uma palavra ou subpalavra que não faz parte do vocabulário;
- (ID=102) - [SEP]: indicativo de espaço entre sentenças ou fim da sentença;
- (ID = 100) - [MASK]: utilizado para mascarar alguns tokens em uma sentença durante a tarefa de *Mask LM* no pré-treino do BERT;
- (ID = 0) - [PAD]: para completar com zeros a representação numérica (vetor) de uma sentença com uma quantidade de tokens menor que o máximo estabelecido, neste caso, 512 tokens.

Após o tokenizador ser treinado e construído seu vocabulário, todo o *corpus* do DepreRedditBR é tokenizado para que as sentenças sejam mapeadas para uma representação numérica correspondente, conforme IDs do vocabulário. Também é gerado, para cada sentença, um vetor chamado de *attention mask*. A função dessa representação vetorial é indicar os tokens que "merecem atenção"quando forem recebidos no pré-treino do modelo. Tomando como exemplo a sentença da Tabela 1, o *attention mask* seria da seguinte forma: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0 ....0]. Ou seja, os uns (1) indicam que nesta posição na representação de IDs numéricos existe um token de uma palavra ou subpalavra ou ainda, um token de controle como o [CLS] ou [SEP.], já os zeros (0) representa que nessa posição da representação de IDs numéricos existem um token de controle [PAD].

Ao final da etapa de tokenização, obtem-se como produto resultante o tokenizador em si, seu vocabulário, o *corpus* tokenizado e um arquivo de configuração. Esses elementos são pré-requisitos para a etapa do pré-treino (etapa 4).

4.4 Pré-treino do modelo DepreBERTBR

O pré-treinamento de um modelo de linguagem como o BERT tem como objetivo treinar o modelo para aprender sobre um contexto, um domínio do problema em que está inserido

e o idioma padrão, considerando o *corpus* utilizado na etapa de tokenização. Ao final do treinamento, o modelo pré-treinado obtido pode ser utilizado em tarefas diversas de PLN como, por exemplo, classificação de texto. O contexto aprendido e as representações numéricas das palavras (*embbedings*) criadas no pré-treinamento do modelo permitem que o ele possa ser retreinado para adaptar-se às nuances de um conjunto de dados específico, conforme tarefa alvo (e.g., classificação, reconhecimento de entidades).

Na etapa de pré-treino (etapa 4 da Figura 10), o modelo DepreBERTBR recebe então como parâmetros o tokenizador pré-treinado, juntamente com o vocabulário criado a partir do *corpus* DepreRedditBR. O *corpus* DepreRedditBR, depois de tokenizado é utilizado pelo DepreBERTBR na sua atividade de pré-treino. Cabe destacar que o pré treino de um modelo de linguagem exige bastante poder computacional, diante da imensa quantidade de dados textuais processada e representações numéricas que devem ser mapeadas a partir do processamento do vocabulário e das sentenças. Por questão de limite orçamentário atribuído ao trabalho para contratação de serviço na nuvem, o DepreBERTBR foi instanciado na versão BERT Base, versão mais leve do BERT configurada para execução com 12 camadas de transformers, um vetor com 768 dimensões e 12 cabeças de autoatenção *self-attention heads*, totalizando 110 Milhões de parâmetros. O pré-treino do DepreBERTBR foi realizado apenas com a tarefa *Masked Language Modeling - Masked LM* do BERT, explicada na Seção 2.8.

4.4.1 Definição de hiperparâmetros

Antes de iniciar o pré-treino do modelo derivado do BERT, o *corpus* DepreRedditBR é dividido em um conjunto de dados de treinamento e um de avaliação. Isso porque durante o pré-treinamento o modelo usa o *corpus* de treino para o seu aprendizado (80%), enquanto que o *corpus* de avaliação (20%) é usado para verificar como o modelo está aprendendo, no caso da tarefa de Masked LM, avalia o quanto o modelo está fazendo previsões corretas dos tokens mascarados. Além disso, alguns hiperparâmetros são configurados antes de inicializar o pré-treino. Hiperparâmetros são configurações que são ajustadas para controlar como o processo de pré-treinamento vai ser comportar. A Tabela 4 apresenta as configurações de hiperparâmetros adotadas para o treinamento do DepreBERTBR. O modelo proposto foi treinado com 10 épocas e executou 221 mil passos. Ao definir um número de épocas igual a 10, implica dizer que o modelo analisou todo o conjunto de dados 10 vezes, com o intuito de aprimorar o seu conhecimento sobre os dados. Nos primeiros 200 mil passos, a taxa de aprendizado (*Learning rate*) foi de 5e-5, valor padrão para o BERT na biblioteca Hugging Face[6]. Para os 21 mil passos restantes, a taxa de aprendizado foi 1e-4. A mudança na taxa de aprendizado teve como finalidade melhorar o desempenho do DepreBERTBR durante o pré-treinamento, aumentando seu aprendizado. A estratégia de avaliação (*Evaluation strategy*) do modelo durante o pré-treino é definida como *steps*, ocorrendo a cada 5000 mil passos, valor definido no parâmetro registro

[6] https://huggingface.co/

Tabela 4 – Configuração dos Hiperparâmentros do pré-treino do DepreBERTBR.

Parâmetro		**Valor**
Otimizador	Optimizer	AdamW
Taxa de aprendizado	Learning rate	5e-5 & 1e-4
Tamanho do lote de treinamento	Train batch size	24
Tamanho do lote de avaliação	Eval batch size	24
Época	Epoch	10
Estratégia de avaliação	Evaluation strategy	steps
Intervalo de Registro do passo	Logging steps	5000
Salvar passos	Save steps	5000

Fonte: Elaborada pelo autor.

de etapas (*Logging steps*). Depois da avaliação, o modelo era salvo em *checkpoints*, uma espécie de modelo temporário com as configurações e pesos aprendidos até determinado ponto de progresso no pré-treinamento. No caso de haver alguma interrupção durante o pré-treino, não é preciso recomeçá-lo desde o início, pois basta retomar o pré-treino a partir de um *checkpoint* selecionado.

4.4.2 Ambiente Computacional para o treinamento

O código-fonte para pré-treinamento do modelo DepreBERTBR foi desenvolvido no ambiente Google Colaboratory Pro+, utilizando uma GPU NVIDIA Ampere A100 Tensor Core e levou 4 dias para sua conclusão. Para realização das tarefas de pré-processamento de dados e de geração do modelo de linguagem, utilizou-se a biblioteca Hugging Face. Com a conclusão da etapa de pré-treinamento, o modelo pré-treinado DepreBERTBR se tornou apto a ser salvo fisicamente, podendo ser instanciado e ajustado para tarefas específicas.

4.5 Ajuste fino

Após o pré-treinamento, o modelo passou para a etapa de ajuste fino (etapa 6 da Figura 10). No ajuste fino ocorre o AT. O modelo DepreBERTBR, então, é ajustado para uma tarefa de PLN em particular como, no caso deste trabalho, a classificação de texto. Dessa forma, foram realizados os ajustes para o DepreBERTBR a partir de um conjunto de dados criado por Sampath e Durairaj (2022). Apenas para reforçar, este conjunto de dados norteia o problema de classificação que fará uso do conhecimento adquirido pelo DepreBERTBR, sendo constituído por postagens do Reddit e rotulado por especialistas de domínio em uma das 3 (três) classes: Sem depressão, Depressão moderada e Depressão grave. Originalmente, esse conjunto de dados foi criado no idioma inglês, por isso foi necessário realizar a tradução dele para o português brasileiro usando a API do Google Translate, ou seja, o mesmo procedimento utilizado para traduzir os *corpus* em inglês incorporados ao DepreRedditBR explicado na Seção 4.1. Após a tradução do conjunto de dados, foi necessário realizar ações de pré-processamento dos da-

Tabela 5 – Amostra do *corpus* rotulado.

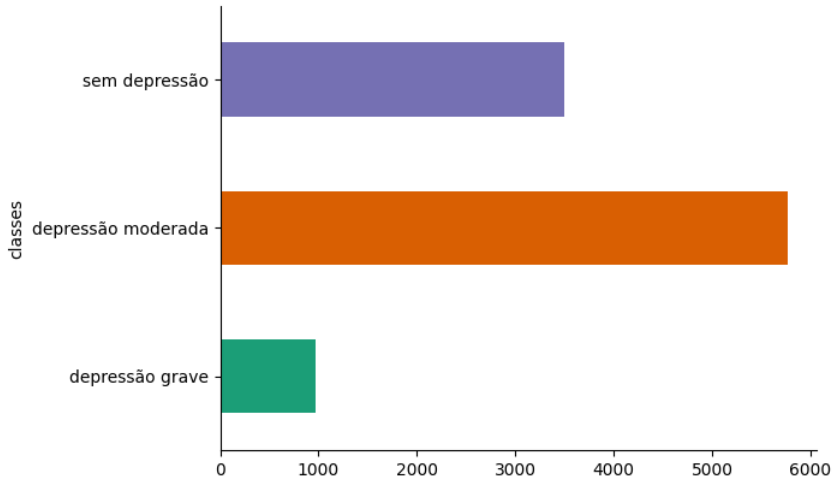
Classe	Texto	Rótulo
Sem depressão	Alguém quer só conversar?	0
Depressão moderada	Estou em uma festa de ano novo e quero chorar, estou tendo um ataque de pânico em motivo algum, não quero mais estar aqui	1
Depressão grave	Eu não quero morrer, só quero parar de viver. Isso faz sentido? Deus, eu odeio minha vida	2

Fonte: Elaborado pelo autor.

dos, sendo aplicados os mesmos ajustes que o ocorreram no *corpus* DepreRedditBR (Ver Seção 4.2). Após o pré-processamento dos dados, o conjunto de dados rotulado apresentou um total de 10.230 postagens. A Tabela 5 mostra um fragmento do *corpus* pré-processado usado para treino e teste do classificador. A coluna *classes* representa o rótulo da postagem referente ao conteúdo da coluna "texto". No caso do ajuste fino alvo deste trabalho, o modelo DepreBERTBR é instanciado como um classificador e recebe os rótulos das classes, assim como em um problema de classificação supervisionada convencional. Esses rótulos também são convertidos para números para que sejam usados pelo DepreBERTBR. Assim as (três) classes Sem depressão, Depressão moderada"e Depressão Grave foram convertidas para 0, 1 e 2, respectivamente.

O gráfico apresentado na Figura 11 mostra a distribuição de frequência das classes existentes no *corpus* de treinamento e teste do classificador. A classe Sem depressão possui 3.495 (34,16%) postagens , a classe Depressão moderada apresenta a maior quantidade de dados com 5.768 (56,39%) postagens , já a classe Depressão grave é a que tem menor quantitativo de dados com 967 (9,45 %) postangens.

Figura 11 – Distribuição das classes do *corpus* rotulado.



Fonte: Elaborado pelo autor.

A criação e avaliação do classificador foram realizadas usando a técnica de validação cruzada estratificada com k=10. A implementação da validação cruzada estratificada contribui para um balanceamento das classes em cada *fold*. A cada iteração (fold), cada partição do *corpus* da tarefa de classificação é tokenizado utilizando o tokenizador do DepreBERTBR apresen-

Tabela 6 – Configuração dos Hiperparâmentros do ajuste fino para tarefa de classificação em 2 cenários.

Cenário com 2 épocas			**Cenário com 10 épocas**		
Parâmetro		**Valor**	**Parâmetro**		**Valor**
Otimizador	Optimizer	AdamW	Otimizador	Optimizer	AdamW
Taxa de aprendizado	Learning rate	5e-5	Taxa de aprendizado	Learning rate	5e-5
Tamanho do lote de treinamento	Train batch size	24	Tamanho do lote de treinamento	Train batch size	32
Tamanho do lote de avaliação	Eval batch size	24	Tamanho do lote de avaliação	Eval batch size	32
Época	Epoch	2	Época	Epoch	10
Estratégia de avaliação	Evaluation strategy	epoch	Estratégia de avaliação	Evaluation strategy	epoch

Fonte: Elaborada pelo autor.

tado na Seção 4.3. o modelo DepreBERTBR é treinado para a tarefa de classificação de texto com as três classes já mencionadas. Alguns hiperparâmetros são configurados antes do início do treinamento, como pode ser observado na Tabela 6. Neste caso, a estratégia de avaliação foi configurada para ser avaliada a cada época, haja visto que o *corpus* com a quantidade de 10.230 postagens apresenta poucos passos. Dessa forma, o modelo pode ver todas as amostras de treinamento para depois realizar a avaliação.

5 EXPERIMENTOS E RESULTADOS

Este capítulo apresenta a avaliação da abordagem proposta baseada na construção do modelo DepreBERTBR. Depois de realizado o ajuste fino para a tarefa de classificação de textos apresentado na Seção 4.5, dois experimentos foram realizados com o intuito de responder às questões de pesquisa definidas no Capítulo 1. Para isso, o primeiro experimento busca avaliar o modelo de linguagem DepreBERTBR implementado neste trabalho, comparando-o com um modelo de domínio geral no idioma português. O segundo experimento busca avaliar o modelo DepreBERTBR com respeito a outro modelo construído considerando o idioma português brasileiro e que possui parte de seus dados no contexto de saúde mental. Os resultados obtidos nos experimentos são discutidos à luz das questões de pesquisa.

5.1 Configuração básica

Os modelos foram construídos para o idioma português brasileiro. A tarefa de classificação é multiclasse e provê a predição de 3 classes de acordo com o nível de depressão: Sem depressão, Depressão moderada, ou Depressão grave. O conjunto de dados utilizado nos experimentos foi criado por Sampath e Durairaj (2022) e traduzido do idioma inglês para o idioma português brasileiro usando a API do Google Translate.

Os experimentos foram divididos em dois cenários. Em ambos os cenários os modelo DepreBERTBR e os modelos BERTimbau e BERTabaporu foram ajustados para a tarefa de classificação utilizando o conjunto de dados rotulado. No primeiro cenário, os modelos foram configurados para serem treinados com 2 épocas. Já no segundo cenário, os modelos foram configurados para serem treinados com 10 épocas, padrão normalmente empregado nos treinamentos dos modelos comparados na literatura. A mudança na quantidade de épocas durante o treinamento tem como intuito analisar se, a partir de uma calibração maior do conhecimento sobre os dados, os modelos apresentarão resultados de predição mais precisos. A Tabela 6 apresentada na Seção 4.5 mostra as configurações dos hiperparâmentros utilizados nos experimentos para ambos os cenários.

Em ambos os cenários os modelos foram instanciados utilizando a versão mais leve (Base) de cada modelo. Para cada cenário de execução foi realizada uma validação cruzada estratificada com 10 dobras. Seguindo os trabalhos relacionados a esta pesquisa apresentados no Capítulo 3, as métricas de avaliação utilizadas para comparar o desempenho dos modelos na tarefa de classificação foram (PAES; VIANNA; RODRIGUES, 2023) : f1, revocação e precisão.

Tabela 7 – Comparativo das métricas de avaliação dos Experimentos.

Cenário	DepreBERTBR			BERTimbau			BERTabaporu		
	Precisão	**Revocação**	**F1**	**Precisão**	**Revocação**	**F1**	**Precisão**	**Revocação**	**F1**
2 épocas	**0,90**	**0,88**	**0,89**	0,88	0,85	0,86	0,90	0,88	0,89
10 épocas	0,89	0,86	0,87	**0,91**	**0,90**	**0,90**	**0,90**	**0,87**	**0,88**

Fonte:Elaborada pelo autor.

5.1.1 Experimento 1

O objetivo deste experimento é analisar os modelos pré-treinados DepreBERTBr e BERTimbau com a intenção de comparar se um modelo treinado com dados de domínio específico tem desempenho melhor em relação a um modelo treinado com dados de domínio geral, com respeito à tarefa de classificação de postagens com possível teor depressivo do Reddit.

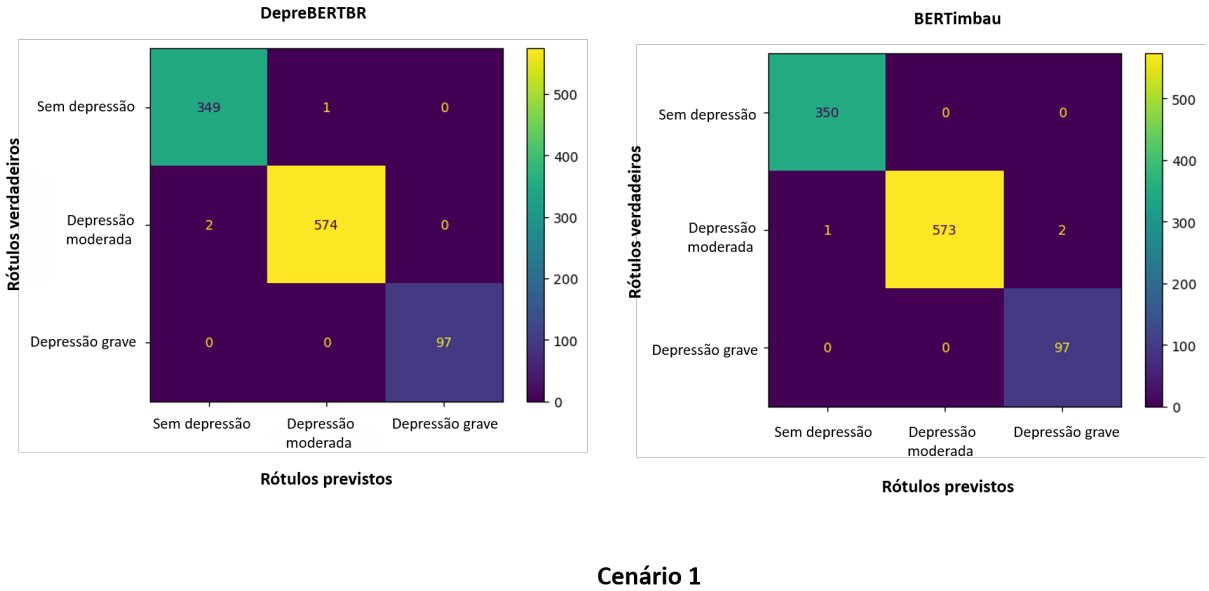
A Tabela 7 apresenta os resultados médios das métricas de avaliação considerando os dois cenários de configuração. No cenário com 2 épocas, o DepreBERTBR superou o desempenho do BERTimbau em todas as métricas de avaliação. Entretanto, no cenário com 10 épocas, o BERTimbau apresentou melhor resultado, com uma leve redução do desempenho do DepreBERTBR. O aumento na quantidade de épocas do Cenário 1 para o Cenário 2 demonstra que, para o BERTimbau, quanto mais o modelo percorre os dados, mais ele aprende e consegue executar a tarefa de classificação de texto com mais eficácia, mesmo que em um domínio de aplicação mais específico. Isso porque, mesmo o BERTimbau não sendo um modelo de linguagem específico para o domínio da depressão, ele foi treinado com um *corpus* imensamente maior quando comparado ao *corpus* do DepreBERTBR, como pode ser observado na Tabela 8. No contexto dos modelos de linguagem, quanto mais dados o modelo conhece durante seu treinamento, mais ele aprende sobre o contexto e as relações entre as palavras. Mesmo assim, apesar de ter apresentando uma perda leve de desempenho quando treinado com 10 épocas, o DepreBERTBR ainda conseguiu obter bons resultados na classificação de textos considerando os diferentes graus de depressão.

As matrizes de confusão[1] apresentadas na Figura 12 e a Figura 13 mostram como o DepreBERTBR e o BERTimbau realizaram a classificação das postagens em cada cenário do experimento, considerando as 3 (três) classes alvo: "Sem depressão"(350), "Depressão moderada"(576) e "Depressão grave"(97). No Cenário 1, podemos observar que o DepreBERTBR e o BERTimbau apresentaram um desempenho equilibrado, errando apenas 3 (três) predições cada. O BERTimbau se mostrou eficiente em classificar corretamente todas as instâncias das classes "sem depressão"e "depressão grave". Porém, classifica incorretamente 3 instâncias da classe "depressão moderada"como "sem depressão"e "depressão grave". O DepreBERTBR é efetivo na classificação correta de todas as instâncias da classe "depressão grave", entretanto, classifica incorretamente apenas 2 instâncias da classe "depressão moderada"e 1 instância da classe "sem depressão".

[1] Matriz de confusão da última rotação da validação cruzada.

No Cenário 2, o DepreBERTBR apresentou quatro falsos negativos para a classe "sem depressão", classificando-os como "depressão moderada", e um falso positivo para a classe "sem depressão"pertencente à classe "depressão moderada". Com um maior número de épocas, o BERTimbau melhorou seu desempenho na classificação das instâncias da classe "depressão moderada", apresentando apenas dois enganos ao classificar instâncias da classe "sem depressão"como "depressão moderada". Em resposta à QP1, percebe-se que o uso de modelos de linguagem apresentam um bom resultado para classificar postagens de redes sociais no idioma português do Brasil, conforme grau de depressão.

Figura 12 – Matriz de confusão do Cenário 1 no Experimento 1.



Cenário 1

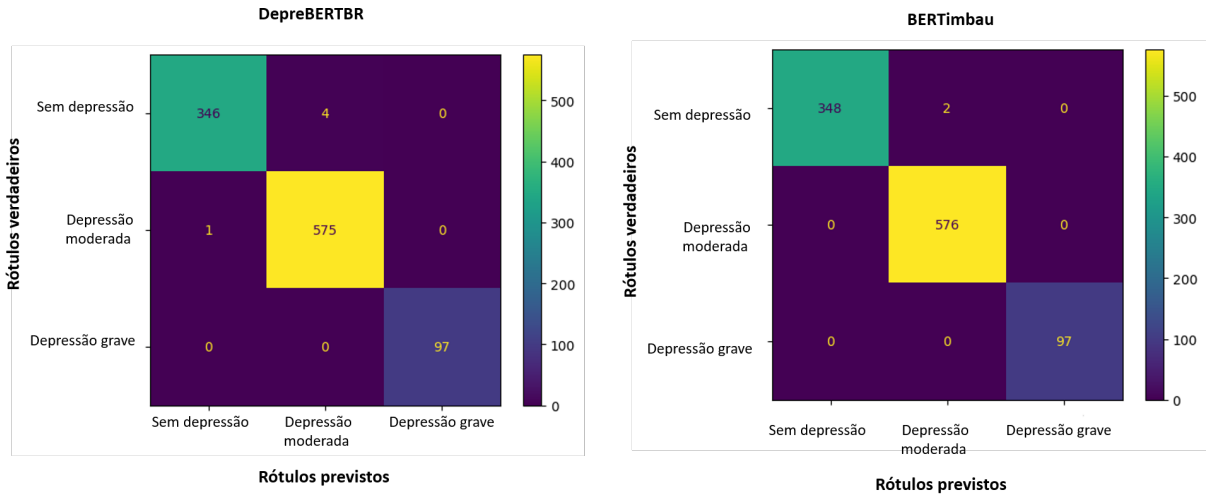
Fonte: Elaborado pelo autor.

5.1.2 Experimento 2

O objetivo do Experimento 2 é analisar os modelos pré-treinados DepreBERTBr e BERTabaporu, ambos na versão Base, com a intenção de comparar se um modelo treinado com dados de domínio específico tem desempenho melhor em relação a um modelo treinado com dados de domínio geral juntamente com dados de transtornos mentais no idioma português, com respeito à tarefa de classificação de postagens com teor de depressão do Reddit.

A Tabela 7 mostra o resultado das métricas Precisão, Revocação e F1 dos cenários do experimento 2. No Cenário com 2 épocas, o DepreBERTBR apresentou exatamente o mesmo resultado em relação ao BERTabaporu em todas as métricas. No Cenário com 10 épocas, o BERTabaporu apresentou uma diferença numérica muito pequena em relação ao DepreBERTBR (0,01) em termos de F1. Neste comparativo, o DepreBERTBR mostrou que permanece competitivo para o número de épocas igual a 10, sabendo-se que o BERTabaporu realizou o treinamento em um *corpus* muito maior. No caso do BERTabaporu, os dados utilizados para o seu pré-treino

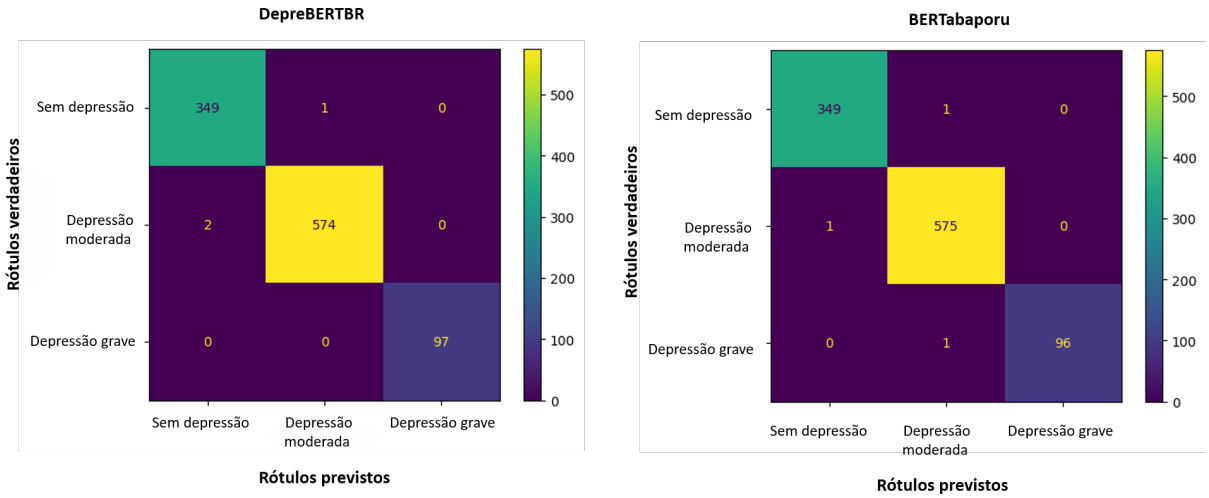
Figura 13 – Matriz de confusão do Cenário 2 no Experimento 1.



Cenário 2

Fonte: Elaborado pelo autor.

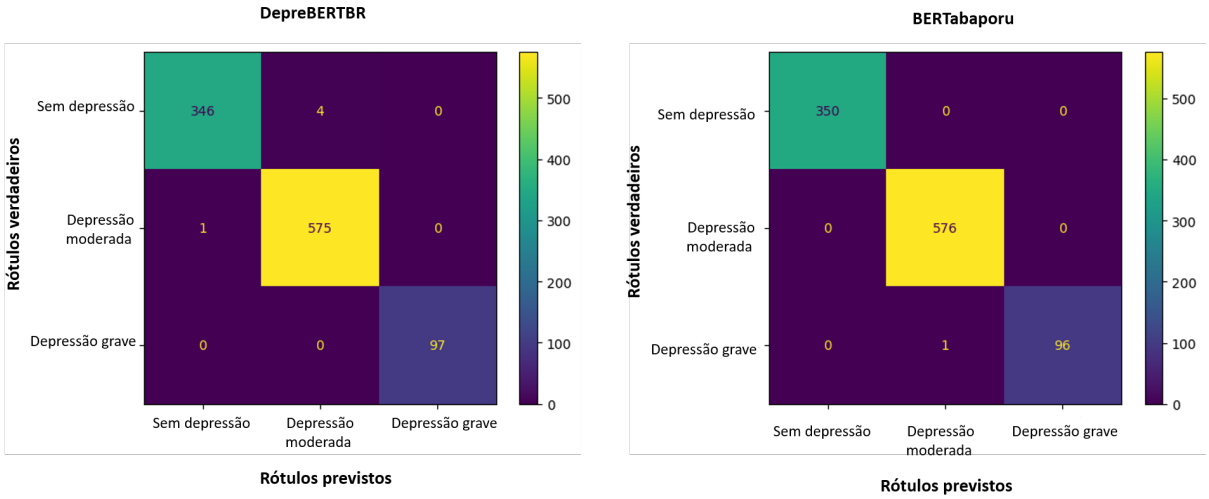
Figura 14 – Matriz de confusão do Cenário 1 no Experimento 2.



Cenário 1

Fonte: Elaborado pelo autor.

Figura 15 – Matriz de confusão do Cenário 2 no Experimento 2.



Cenário 2

Fonte: Elaborado pelo autor.

também contavam com dados sobre saúde mental a partir de postagens no Twitter como pode ser observado na Tabela 8.

A matriz de confusão [2] apresentada na Figura 14 mostra que no Cenário 1 o DepreBERTBR acerta a classificação de todas as instâncias da classe "depressão grave"e classifica incorretamente apenas 3 instâncias: 2 falsos positivos da classe "sem depressão", que percentem à classe "depressão moderada", e 1 falso positivo da classe "depressão moderada", que pertence à classe "sem depressão". O BERTabaporu também só apresentou 3 classificações incorretas, dois falsos positivos da classe "depressão moderada", cada uma oriunda das classes "sem depressão"e e "depressão grave", e 1 falso positivo da classe "sem depressão", que na verdade pertence à classe "depressão moderada".

Já no Cenário 2 a Figura 15 mostra um cenário em que o BERTAbaporu conseguiu classificar corretamente todos os exemplos das classes "sem depressão"e "depressão moderada", errando apenas 1 previsão da classe "depressão grave"como "depressão moderada". O DepreBERTBR então, incrementou o número de classificações incorretas da classe "sem depressão"como sendo "depressão moderada".

5.2 Considerações sobre a avaliação

A Tabela 8 mostra um resumo das características dos dados utilizados para o pré-treinamento dos modelos DepreBERTBR, BERTimbau e BERTabaporu. É importante destacar a diferença considerável da quantidade de textos utilizada para o pré-treino de cada modelo.

[2] Matriz de confusão da última rotação da validação cruzada.

Tabela 8 – Comparativo das características dos modelos pré-treinados

Modelo	**Domínio**	**Origem**	**Tamanho do corpus**	**Perc. (%) de 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

Fonte: Elaborada pelo autor.

Tanto o BERTimbau quanto o BERTabaporu foram treinados com milhões de sentenças, enquanto o DepreBERTBR utilizou milhares. No entanto, os experimentos revelam que modelo desenvolvido nesta dissertação mostrou-se muito competitivo no comparativo com o BERTimbau e BERTabaporu para a tarefa de classificação de postagens no idioma português brasileiro, considerando o grau de depressão, apresentando um resultado de avaliação equitativo Dessa forma, em resposta à QP2, fica evidente que há um potencial para classificar postagens utilizando um modelo pré-treinado para um domínio e idioma específicos, particularmente, o domínio da depressão no idioma português brasileiro. Ainda, o modelo proposto nesta dissertação consegue alcançar um resultado compatível com os modelos comparados, com uma massa de dados consideravelmente menor.

A Tabela 9 mostra um comparativo entre o trabalho desenvolvido nesta pesquisa e os trabalhos relacionados aprensentados no Capítulo 3. O DepreBERTBR realizou um pré-treinamento de um modelo de linguagem específico para o domínio da depressão no idioma português brasileiro com o objetivo de ajustá-lo para classificar postagens do Reddit considerando o grau de depressão. Os trabalhos de Ji et al. (2022) e Poświata e Perełkiewicz (2022) também utilizaram postagens do Reddit com teor depressivo, entretanto, esses modelos foram pré-treinados para o idioma inglês. A Revisão Sistemática da Literatura desenvolvida por (HERCULANO et al., 2022) mostrou que a maioria das pesquisas para detectar depressão é no idioma inglês, evidenciando uma lacuna no que se refere a trabalhos no idioma português para detectar depressão.

Para o treinamento de modelos de linguagem é preciso uma enorme quantidade de textos. Para o treinamento do DepreBERTBR foi necessário construir um *corpus* com postagens com conteúdo depressivo. Dessa forma, o *corpus* DepreRedditBR foi criado a partir de postagens nas subcomunidade (subreddits) do Reddit com teor depressivo na língua portuguesa do Brasil, em conjunto com *corpus* traduzido do inglês para o português. O trabalho de Santos, Oliveira e Paraboni (2023) também criou um *corpus* no idioma português brasileiro, no entanto, utilizou como fonte de dados a rede social Twitter (atualmente chamada de X), filtrando usuários que já haviam sido diagnosticado com depressão e/ou transtorno de ansiedade.

Em relação aos modelos pré-treinados em português,o BERTimbau e o BERTabaporu foram treinados com dados de domínio geral. Todavia, o BERTabaporu incluiu também no seu treinamento, dados sobre saúde mental a partir de postagens no Twitter, conforme mostra a

Tabela 8. Por outro lado, o DepreBERTBR focou apenas em um único domínio no treinamento, a depressão.

Tabela 9 – Comparativo entre os trabalhos relacionados e o DepreBERTBR.

Trabalho	**Domínio de dados**	**Realizou pré-treino**	**Fine Tuning**	**Idioma**	**Corpus**	**Contribuições**
MentalBert	Transtornos mentais	Sim	Sim	Inglês	Produziu	Desenvolveu dois modelos pré-treinados baseados no BERT e no RoBERTa para o domínio dos transtornos mentais no idioma inglês
DepRoberta	Depressão	Sim	Sim	Inglês	Utilizou existente	Desenvolveu um modelo pré-treinado em inglês baseado no RoBERTa, Classificou as postagens em nível de acordo com o nível de depressão
BERTimbau	Geral	Sim	Sim	Português	Utilizou existente	Desenvolveu um modelo pré-treinado em português baseado no BERT com o domínio de dados geral
SetembroBr	Depressão e Ansiedade	Não	Não	Português	Produziu	Construiu um corpus nos domínios da depressão e ansiedade em português
BERTabaporu	Geral com dados sobre saúde mental	Sim	Sim	Português	Produziu	Desenvolveu um modelo pré-treinado em português baseado no BERT com foco no gênero de textos das redes sociais
DepreBERTBR	**Depressão**	**Sim**	**Sim**	**Português**	**Produziu**	**Desenvolveu um modelo de linguagem pré-treinado no domínio específico da depressão no idioma português brasileiro**

Fonte: Elaborada pelo Autor.

6 CONSIDERAÇÕES FINAIS

A depressão é um transtorno que vem incapacitando a população em nível mundial e tem sido motivo de alerta pela Organização Mundial de Saúde (OMS). A depressão pode causar baixa auto estima, tristeza, sentimentos de culpa e, em casos mais graves, pode levar o indivíduo à morte. Pesquisadores têm utilizado postagens em redes sociais com o intuito de detectar indícios de depressão. Esta dissertação de mestrado apresentou o DepreBERTBR, um modelo de linguagem pré-treinado para o domínio específico da depressão no idioma português brasileiro. A abordagem para a construção do DepreBERTBR foi desenvolvida a partir do modelo pré-treinado BERT e utilizou um *corpus* com postagens no idioma português brasileiro. As postagens com teor depressivo foram coletadas a partir de subcomunidades do Reddit utilizando uma ferramenta de extração. As postagens pertencentes a outros conjuntos de dados no idioma inglês, foram incorporadas ao *corpus* com o suporte de uma ferramenta de tradução de postagens do idioma inglês pra o português brasileiro. A partir dos dados obtidos, um *corpus* denominado DepreReddit foi gerado. O DepreBERTBR foi pré-treinado utilizando o *corpus* DepreReddit e ajustado para a tarefa de classificação de texto, particularmente, textos de postagens no Reddit com teor depressivo. A classificação dos textos das postagens considerou três graus de depressão: "Sem depressão", "Depressão moderada"e "Depressão grave". A avaliação experimental realizada demonstra que o modelo desenvolvido nesta dissertação é bastante competitivo em comparação a outros modelos no idioma português do Brasil para tarefas de classificação, especialmente, para detectar sinais de depressão.

6.1 Principais contribuições

As principais contribuições deste trabalho são elencadas a seguir:

- Criação de um *Corpus* denominado DepreRedditBR, com 509.189 postagens coletadas do Reddit com teor depressivo no idioma português brasileiro;
- Um Modelo de linguagem pré-treinado denominado DepreBERTBR para o domínio da depressão no idioma português brasileiro que pode ser ajustado para tarefas de PLN, avaliado, em particular, em um problema multiclasse de classificação de textos com ou sem indícios de depressão;
- Um Classificador treinado com base no DepreBERTBR para identificar postagens no idioma português brasileiro com ou sem indícios de depressão, considerando três classes: "Sem depressão", "Depressão moderada", "Depressão grave".

6.2 Trabalhos Futuros

Como trabalhos futuros, são apontados os seguintes direcionamentos:

- Incrementar o *corpus* DepreRedditBR com mais postagens no idioma português brasileiro no domínio da depressão para incrementar o pré-treino do modelo DepreBERTBR;
- Pré-treinar o modelo DepreBERTBR utilizando como base a versão maior do BERT, chamada de BERT Large;
- Realizar o ajuste fino do modelo DepreBERTBR para a tarefa de classificação de textos no idioma português brasileiro no domínio da depressão, de acordo o Inventário de Depressão de Beck que considera quatro níveis de depressão: "Sem depressão", "Depressão leve", "Depressão moderada"e "Depressão grave."

REFERÊNCIAS BIBLIOGRÁFICAS

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020. Citado 2 vezes nas páginas 22 e 23.

ALSENTZER, E. et al. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019. Citado 2 vezes nas páginas 17 e 39.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. [S.l.]: American psychiatric association Washington, DC, 2013. v. 5. Citado 2 vezes nas páginas 15 e 19.

AZAM, F. et al. Identifying depression among twitter users using sentiment analysis. In: IEEE. *2021 international conference on artificial intelligence (ICAI)*. [S.l.], 2021. p. 44–49. Citado na página 47.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. Citado na página 27.

BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. Métodos para análise de sentimentos em mídias sociais. *Sociedade Brasileira de Computação*, 2015. Citado 2 vezes nas páginas 25 e 26.

BENGIO, Y.; DUCHARME, R.; VINCENT, P. A neural probabilistic language model. *Advances in neural information processing systems*, v. 13, 2000. Citado na página 27.

BROWN, T. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020. Citado 2 vezes nas páginas 29 e 32.

CACHEDA, F. et al. Early detection of depression: social network analysis and random forest techniques. *Journal of medical Internet research*, JMIR Publications Inc., Toronto, Canada, v. 21, n. 6, p. e12554, 2019. Citado 2 vezes nas páginas 15 e 16.

CAÑETE, J. et al. Spanish pre-trained bert model and evaluation data. *arXiv preprint arXiv:2308.02976*, 2023. Citado na página 17.

CASELI, H. d. M.; NUNES, M. d. G. V. *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*. 2a. ed. [S.l.]: BPLN, 2023. ISBN 978-65-00-95750-1. Citado 6 vezes nas páginas 26, 27, 28, 29, 30 e 32.

CHOUDHURY, M. D. et al. Predicting depression via social media. In: *Seventh international AAAI conference on weblogs and social media*. [S.l.: s.n.], 2013. p. 128–137. Citado 2 vezes nas páginas 16 e 20.

CHOWDHERY, A. et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, v. 24, n. 240, p. 1–113, 2023. Citado na página 32.

COPPERSMITH, G. et al. Clpsych 2015 shared task: Depression and ptsd on twitter. In: *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*. [S.l.: s.n.], 2015. p. 31–39. Citado na página 39.

COSTA, P. B. et al. Bertabaporu: assessing a genre-specific language model for portuguese nlp. In: *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*. [S.l.: s.n.], 2023. p. 217–223. Citado 2 vezes nas páginas 42 e 43.

CUNHA, R. V. d.; BASTOS, G. A. N.; DUCA, G. F. D. Prevalência de depressão e fatores associados em comunidade de baixa renda de porto alegre, rio grande do sul. *Revista Brasileira de Epidemiologia*, SciELO Public Health, v. 15, p. 346–354, 2012. Citado na página 15.

DAI, A. M.; LE, Q. V. Semi-supervised sequence learning. *Advances in neural information processing systems*, v. 28, 2015. Citado na página 28.

DENG, S.; SINHA, A. P.; ZHAO, H. Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*, Elsevier, v. 94, p. 65–76, 2017. Citado na página 16.

DEVLIN, J. et al. 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*. NAACL, 2019. p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423>. Citado 5 vezes nas páginas 17, 29, 30, 33 e 35.

DUQUE, J. W. G.; RAYMUNDO, A. L.; NETO, P. F. An application of big data for twitter depressive sentence classification. *H-TEC humanities and technology magazine*, v. 2, n. 1, p. 82–95, 2018. Citado na página 16.

ESTRELA, P. et al. Análise de sentimentos em postagens do reddit no intercurso da pandemia de covid-19. *Submetido à Revista Principia*, 2024. Citado na página 46.

FILHO, J. A. W. et al. The brwac corpus: a new open resource for brazilian portuguese. In: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. [S.l.: s.n.], 2018. Citado na página 36.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>. Citado na página 23.

GORENSTEIN, C.; ANDRADE, L. Inventário de depressão de beck: propriedades psicométricas da versão em português. *Rev psiq clin*, v. 25, n. 5, p. 245–50, 1998. Citado na página 20.

GOVINDASAMY, K. A.; PALANICHAMY, N. Depression detection using machine learning techniques on twitter data. In: IEEE. *2021 5th international conference on intelligent computing and control systems (ICICCS)*. [S.l.], 2021. p. 960–966. Citado na página 17.

HAN, J.; KAMBER, M.; PEI, J. Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*, 2012. Citado 2 vezes nas páginas 22 e 23.

HARRINGTON, P. *Machine learning in action*. [S.l.]: Simon and Schuster, 2012. Citado 2 vezes nas páginas 22 e 23.

HERCULANO, A. et al. Detecting signs of mental disorders on social networks: a systematic literature review. *DATA ANALYTICS 2022*, p. 55–61, 2022. Citado 4 vezes nas páginas 20, 43, 47 e 60.

JI, S. et al. MentalBERT: Publicly available pretrained language models for mental healthcare. In: CALZOLARI, N. et al. (Ed.). *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022. p. 7184–7190. Disponível em: <https://aclanthology.org/2022.lrec-1.778>. Citado 6 vezes nas páginas 17, 21, 22, 38, 43 e 60.

KAYALVIZHI, S. et al. Findings of the shared task on detecting signs of depression from social media. In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. [S.l.: s.n.], 2022. p. 331–338. Citado na página 39.

KUDO, T.; RICHARDSON, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. Citado na página 30.

LEE, J. et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Oxford University Press, v. 36, n. 4, p. 1234–1240, 2020. Citado 2 vezes nas páginas 17 e 39.

LIDDY, E. D. *Natural language processing*. [S.l.]: In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc., 2001. Citado na página 26.

LIGTHART, A.; CATAL, C.; TEKINERDOGAN, B. Systematic reviews in sentiment analysis: a tertiary study. *Artificial intelligence review*, Springer, v. 54, n. 7, p. 4997–5053, 2021. Citado 2 vezes nas páginas 22 e 26.

LIN, H. et al. Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 29, n. 9, p. 1820–1833, 2017. Citado na página 20.

LIN, L. Y. et al. Association between social media use and depression among us young adults. *Depression and anxiety*, Wiley Online Library, v. 33, n. 4, p. 323–331, 2016. Citado na página 16.

LIN, T. et al. A survey of transformers. *AI Open*, v. 3, p. 111–132, 2022. ISSN 2666-6510. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666651022000146>. Citado na página 33.

LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. Citado na página 39.

LOSADA, D. E.; CRESTANI, F. A test collection for research on depression and language use. In: SPRINGER. *International Conference of the Cross-Language Evaluation Forum for European Languages*. [S.l.], 2016. p. 28–39. Citado na página 39.

LOW, D. M. et al. 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*, JMIR Publications Toronto, Canada, v. 22, n. 10, p. e22635, 2020. Citado 2 vezes nas páginas 39 e 48.

MARKOV, A. A.; SCHORR-KON, J. J. *Theory of algorithms*. [S.l.]: Springer, 1962. Citado na página 28.

MARTIN, L. et al. CamemBERT: a tasty French language model. In: JURAFSKY, D. et al. (Ed.). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 7203–7219. Disponível em: <https://aclanthology.org/2020.acl-main.645>. Citado na página 17.

MIGUEL, E. C. et al. *Clínica psiquiátrica: as grandes síndromes psiquiátricas [ampl. e atual.]*. [S.l.]: Manole, 2021. Citado na página 19.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. Citado na página 27.

MITCHELL, T. M. et al. *Machine learning*. [S.l.]: McGraw-hill New York, 1997. Citado na página 22.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, Manole, v. 1, n. 1, p. 32, 2003. Citado na página 22.

NARDI, A. E.; SILVA, A. G. da; QUEVEDO, J. *Tratado de Psiquiatria da Associação Brasileira de Psiquiatria*. [S.l.]: Artmed Editora, 2021. Citado 3 vezes nas páginas 16, 19 e 20.

NASCIMENTO, R. da S. et al. Identificando sinais de comportamento depressivo em redes sociais. In: SBC. *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*. [S.l.], 2018. Citado na página 47.

OLIVAS, E. S. et al. *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques*. [S.l.]: IGI global, 2009. Citado na página 25.

OLIVEIRA, B. S. N. et al. Processamento de linguagem natural via aprendizagem profunda. *Sociedade Brasileira de Computação*, 2022. Citado 5 vezes nas páginas 17, 24, 27, 28 e 31.

ORGANIZATION, W. H. *Depression and other common mental disorders: global health estimates*. [S.l.], 2017. 24 p. p. Citado na página 15.

PAES, A.; VIANNA, D.; RODRIGUES, J. Capítulo 15 modelos de linguagem. 2023. Citado 2 vezes nas páginas 22 e 55.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 22, n. 10, p. 1345–1359, 2009. Citado na página 25.

PÉREZ, A.; PARAPAR, J.; BARREIRO, Á. Automatic depression score estimation with word embedding models. *Artificial Intelligence in Medicine*, Elsevier, v. 132, p. 102380, 2022. Citado na página 20.

PETERS, M. E. et al. Deep contextualized word representations. In: WALKER, M.; JI, H.; STENT, A. (Ed.). *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 2227–2237. Disponível em: <https://aclanthology.org/N18-1202>. Citado na página 29.

PIRES, R. et al. Sabiá: Portuguese large language models. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2023. p. 226–240. Citado na página 32.

PIRINA, I.; ÇÖLTEKIN, Ç. Identifying depression on reddit: The effect of training data. In: *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*. [S.l.: s.n.], 2018. p. 9–12. Citado na página 39.

POLIGNANO, M. et al. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In: *Italian Conference on Computational Linguistics*. [s.n.], 2019. Disponível em: <https://api.semanticscholar.org/CorpusID:204914950>. Citado na página 17.

POŚWIATA, R.; PEREŁKIEWICZ, M. 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*. [S.l.: s.n.], 2022. p. 276–282. Citado 7 vezes nas páginas 17, 22, 39, 40, 43, 48 e 60.

RADFORD, A. et al. Improving language understanding with unsupervised learning. Technical report, OpenAI, 2018. Citado na página 33.

RAO, A. et al. Sentiment analysis on user-generated video, audio and text. In: *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. [S.l.: s.n.], 2021. p. 24–28. Citado na página 25.

REAL, L.; FONSECA, E.; OLIVEIRA, H. G. The assin 2 shared task: a quick overview. In: SPRINGER. *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*. [S.l.], 2020. p. 406–412. Citado na página 36.

RÍSSOLA, E. A.; LOSADA, D. E.; CRESTANI, F. A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, ACM New York, NY, USA, v. 2, n. 2, p. 1–31, 2021. Citado 2 vezes nas páginas 15 e 19.

ROSA, R. L. et al. A knowledge-based recommendation system that includes sentiment analysis and deep learning. *IEEE Transactions on Industrial Informatics*, IEEE, v. 15, n. 4, p. 2124–2135, 2018. Citado na página 16.

SAMPATH, K.; DURAIRAJ, T. Data set creation and empirical analysis for detecting signs of depression from social media postings. In: SPRINGER. *International Conference on Computational Intelligence in Data Science*. [S.l.], 2022. p. 136–151. Citado 3 vezes nas páginas 40, 52 e 55.

SANTOS, D. et al. Harem: An advanced ner evaluation contest for portuguese. In: *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*. [S.l.: s.n.], 2006. Citado na página 37.

SANTOS, F. A. et al. Processamento de linguagem natural em textos de mídias sociais: Fundamentos, ferramentas e aplicações. *Sociedade Brasileira de Computação*, 2022. Citado na página 27.

SANTOS, W. R. d.; OLIVEIRA, R. L. de; PARABONI, I. Setembrobr: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, Springer, p. 1–28, 2023. Citado 3 vezes nas páginas 41, 43 e 60.

SILVA, J. C. da; VIEIRA, R. O. Introdução às redes neurais profundas com python. *Sociedade Brasileira de Computação*, 2022. Citado 2 vezes nas páginas 23 e 24.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*. [S.l.], 2020. p. 403–417. Citado 7 vezes nas páginas 24, 30, 36, 37, 38, 42 e 43.

SPERLING, O. V.; LADEIRA, M. Mining twitter data for signs of depression in brazil. In: *Anais do VII Symposium on Knowledge Discovery, Mining and Learning*. [S.l.]: SBC, 2019. p. 25–32. Citado na página 16.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, v. 27, 2014. Citado na página 27.

TARDELLI, A. V.; DIAS, A. F. d. S.; FRANÇA, J. B. d. S. Introdução à análise de sentimentos com word clouds. *Sociedade Brasileira de Computação*, 2019. Citado na página 25.

TAULLI, T. *Introdução à Inteligência Artificial: Uma abordagem não técnica*. [S.l.]: Novatec Editora, 2020. Citado na página 23.

TLELO-COYOTECATL, I.; ESCALANTE, H. J.; GÓMEZ, M. Montes y. Depression recognition in social media based on symptoms detection. Sociedad Española para el Procesamiento del Lenguaje Natural, 2022. Citado na página 20.

TOUVRON, H. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. Citado na página 32.

TUNSTALL, L.; WERRA, L. V.; WOLF, T. *Natural language processing with transformers*. [S.l.]: "O'Reilly Media, Inc.", 2022. Citado 4 vezes nas páginas 24, 27, 30 e 32.

UBAN, A.-S.; CHULVI, B.; ROSSO, P. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, Elsevier, v. 124, p. 480–494, 2021. Citado na página 16.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado 3 vezes nas páginas 32, 33 e 34.

VEDULA, N.; PARTHASARATHY, S. Emotional and linguistic cues of depression from social media. In: *Proceedings of the 2017 International Conference on Digital Health*. [S.l.: s.n.], 2017. p. 127–136. Citado 2 vezes nas páginas 15 e 16.

WHO. *World Health Organization*. 2022. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders> Last accessed 27 Abril 2024. Citado na página 19.

WHO. *World Health Organization*. 2023. <https://www.who.int/news-room/fact-sheets/detail/depression> Last accessed 27 Abril 2024. Citado na página 15.

WU, Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. Citado 3 vezes nas páginas 30, 34 e 49.

XU, W.; RUDNICKY, A. Can artificial neural networks learn language models? Carnegie Mellon University, 2000. Citado na página 27.

YOSINSKI, J. et al. How transferable are features in deep neural networks? *Advances in neural information processing systems*, v. 27, 2014. Citado na página 24.

ZHAO, W. X. et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. Citado 3 vezes nas páginas 28, 29 e 32.

ZHU, Y. et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 19–27. Citado na página 34.