UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

VANESSA BORBA DE SOUZA

DAC Stacking: Comitê de Redes Profundas para Classificação de Ansiedade, Depressão e Comorbidade

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Profa. Dra. Karin Becker

Porto Alegre
2021

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Souza, Vanessa Borba de

DAC Stacking: Comitê de Redes Profundas para Classificação de Ansiedade, Depressão e Comorbidade / Vanessa Borba de Souza. – Porto Alegre: PPGC da UFRGS, 2021.

151 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2021. Orientador: Karin Becker.

1. Deep Learning. 2. Ensemble. 3. Redes Sociais. 4. Saúde Mental. 5. Word Embeddings. 6. Processamento de Linguagem Natural. I. Becker, Karin. II. Título.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pela saúde e determinação diante dos momentos mais difíceis ao longo desta jornada acadêmica.

Agradeço a minha família, em especial aos meus pais Vilmar e Liane e meu esposo Marcelo pelo apoio, incentivo e compreensão incondicional durante esse período. Aos amigos, em especial Janaína, Márcia e Francielle pelas experiências compartilhadas e incentivo incansável durante todo esse período. Todos foram de extrema importância durante os momentos de incerteza.

Agradeço a minha excelente orientadora, Prof[a]. Dr[a]. Karin Becker, pela oportunidade, conselhos, zelo e ensinamentos durante esses três anos. Ensinamentos os quais foram além do desenvolvimento deste projeto agregando conhecimento para minha vida profissional como um todo. Agradeço também ao Prof. Dr. Jéferson Campos Nobre pela grande contribuição acerca dos conhecimentos em Psicologia, atuando ativamente para melhoria deste trabalho.

Por fim, agradeço a todos do INF-UFRGS, em especial, aos demais professores pelos ensinamentos e contribuições ao longo dessa jornada acadêmica.

RESUMO

A depressão tornou-se um problema de saúde pública mundial, que atinge cerca de 322 milhões de pessoas no mundo, a um custo aproximado de $2,5 trilhões de dólares. A taxa de comorbidade de depressão com ansiedade também é alta, acentuando o quadro clínico de indivíduos deprimidos. A identificação precoce desses distúrbios é um fator crítico para a triagem correta e a decisão apropriada sobre as linhas de tratamento adequadas. O uso das redes sociais como forma de os indivíduos exporem suas dificuldades anonimamente permitiu a ampliação de estudos em saúde mental com o apoio da área computacional. Trabalhos relacionados abordam a identificação automática de condições mentais específicas, com foco na depressão. O presente trabalho expande essas contribuições, propondo um classificador *ensemble* para a identificação automática de depressão, ansiedade e a comorbidade dessas desordens, utilizando um conjunto de dados de autodiagnóstico extraído da rede social Reddit. O uso do método *ensemble* visa superar as dificuldades de lidar com o problema de classificação multirrótulo envolvidas no cenário de comorbidade, onde os padrões distintos podem ser mais difíceis de identificar. O nível mais baixo *ensemble* é composto de classificadores fracos que geram previsões binárias de rótulo único em condições de treinamento específicas e que seguem uma arquitetura de aprendizado profundo. Para o nível *meta-learning*, uma rede neural densa explora esses classificadores fracos como um contexto para se chegar a uma decisão com vários rótulos. Um extenso conjunto de experimentos usando as arquiteturas de aprendizado profundo LSTM, CNN e sua combinação, *word embeddings* e topologias *ensemble* foi desenvolvido. Todos os classificadores fracos e o modelo *ensemble* superaram as linhas de base. Os classificadores binários baseados na CNN obtiveram o melhor desempenho, com medidas F de 0,79 para depressão, 0,78 para ansiedade e 0,78 para comorbidade. A topologia do conjunto que obteve o melhor desempenho (perda de Hamming de 0,27 e taxa de correspondência exata de 0,47) combina classificadores fracos de acordo com três arquiteturas e não inclui classificadores de comorbidade. Também realizamos uma análise qualitativa usando SHAP, e confirmamos que as características influentes estão relacionadas aos sintomas desses distúrbios.

Palavras-chave: Deep Learning. Ensemble. Redes Sociais. Saúde Mental. Word Embeddings. Processamento de Linguagem Natural.

A Deep Learning Ensemble to Classify Anxiety, Depression, and their Comorbidity from Texts of Social Networks

ABSTRACT

Depression has become a worldwide public health problem, affecting approximately 322 million people worldwide. The comorbidity rate of depression with anxiety is also high, accentuating the clinical picture of depressed individuals. The early identification of these disorders is a critical factor for the correct screening and appropriate decision on the proper lines of treatment. The use of social networks as a means for individuals to expose their difficulties anonymously allowed the expansion of studies in mental health with the support of the computational area. Related works address the automatic identification of specific mental conditions, with a focus on depression. The present work expands these contributions by proposing an ensemble classifier for the automatic identification of depression, anxiety, as well as the comorbidity, using a self-diagnosed dataset extracted from Reddit. The use of a stacking ensemble aims to overcome the difficulties of dealing with the multi-class, multi-label classification problem involved in the scenario of comorbidity, where the distinctive patterns may be harder to identify. The stacking is composed of specialized single label binary classifiers that distinguish between specific disorders and control users. A meta-learner explores these weak classifiers as a context for reaching a multi-label, multi-class decision. We developed extensive experiments using alternative architectures (LSTM, CNN, and their combination), word embeddings, and ensemble topologies. All weak classifiers and ensembles outperformed the baselines. The CNN-based binary classifiers achieved the best performance, with f-measures of 0.79 for depression, 0.78 for anxiety, and 0.78 for comorbidity. The ensemble topology that achieved the best performance (Hamming Loss of 0.27 and Exact Match Ratio of 0.47) combines weak classifiers according to three architectures, and do not include comorbidity classifiers. We also performed a qualitative analysis using SHAP, and confirmed the influential features are related to symptoms of these disorders.

Keywords: Deep Learning, Ensemble, Social Network, Mental Health, Word Embeddings, Natural Language Processing.

LISTA DE ABREVIATURAS E SIGLAS

ANEW	Affective Norms for English Words
API	Application Program Interface
AUC	Area Under Curve
BERT	Bidirecional Encoder Representations from Transformers
BoW	Bag of Words
CBOW	Continuous Bag Of Words
CES-D	Center for Epidemiologic Studies Depression Scale
CNN	Convolutional Neural Network
DF	Document Frequence
DNN	Redes Neurais Densas
DSM-5	Diagnostic and Statistical Manual of Mental Disorders V
ELMo	Embeddings from Language Models
EMR	Exact Match Ratio
GBDT	Gradient Boosted Decision Trees
GloVe	Global Vectors
GRU	Gated Recurrent Unit
HL	Hamming Loss
IDF	Inverse Document Frequence
InfoGain	Information Gain
LDA	Linear Discriminant Analysis
LIWC	Linguistic Inquiry andWord Count
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron
OMS	Organização Mundial da Saúde

PCA	Principal Components Analysis
PLN	Processamento de Linguagem Natural
POS	Part of Speech
PP	Pontos Percentuais
ReLU	Unidade Linear Retificada
RNN	Recurrent Neural Network
ROC	Receiving Operating Characteristics
SHAP	SHapley Additive exPlanations
SMHD	Self-reported Mental Health Diagnoses
SVM	Support Vector Machines
Tanh	Tangente Hiperbólica
TDAH	Transtorno de Déficit de Atenção/Hiperatividade
TEPT	Transtorno de Estresse Pós-Traumático
TF	Term Frequence
TOC	Transtorno Obses-sivo Compulsivo
URL	Uniform Resource Locator
VADER	Valence Aware Dictionary and sEntiment Reasoner
WE	Word Embedding

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A depressão é um estado caracterizado pela presença de um humor triste, vazio ou irritável, acompanhado por alterações somáticas e cognitivas que afetam significativamente a capacidade do indivíduo de funcionar, o que prejudica seu desempenho nas tarefas diárias e na vida social (American Psychiatric Association, 2013). Seu impacto incapacitante na sociedade preocupa as entidades de saúde pública, atingindo números alarmantes. Segundo relatório da Organização Mundial da Saúde (OMS), 322 milhões de pessoas sofriam de depressão até 2015, a um custo aproximado de $ 2,5 trilhões de dólares (WHO, 2017).

A ansiedade também está entre as doenças mais incapacitantes da população em nível mundial (WHO, 2017). De acordo com a American Psychiatric Association (2013), a ansiedade inclui transtornos que compartilham características de medo e ansiedade excessivos em vários domínios (filhos, violência, profissão, etc.). Além disso, os indivíduos podem experimentar sintomas físicos que incluem inquietação ou sensação de tensão, dificuldade de concentração, irritabilidade, tensão muscular e distúrbios do sono.

Estudos fornecem evidências da relação de proximidade entre ansiedade e depressão. A taxa de comorbidade[1] dessas desordens é alta, já que cerca de 85% dos pacientes com depressão experienciam sintomas significativos de ansiedade (TILLER, 2013). Já a depressão ocorre em mais de 90% dos pacientes que sofrem de transtornos de ansiedade. Essa comorbidade acentua o quadro clínico do indivíduo deprimido, levando a um maior risco de suicídio, pior funcionamento social e resistência ao tratamento (HIRSCHFELD, 2001). O impacto imposto pela depressão na sociedade requer estratégias de intervenção preventiva, particularmente no que diz respeito à identificação precoce dos sintomas (HAMILTON, 1967; RADLOFF, 1977). Esse diagnóstico precoce pode levar à intervenção antecipada; uma redução nos impactos negativos do distúrbio (HALFIN, 2007); a melhores programas de apoio; e a uma melhor compreensão da relação entre condições graves e prevalentes (por exemplo, depressão e ansiedade).

A tarefa de diagnosticar um indivíduo que sofre de uma desordem mental envolve diferentes habilidades, que vão desde a percepção e interpretação dos relatos do paciente, até a distinção sutil de sintomas entre patologias que apresentam comportamentos comuns, tais como a ansiedade e a depressão. Entretanto, essa tarefa é complexa mesmo para os profissionais de saúde, quando não especializados em psicologia ou psiquiatria.

[1] Ocorrência concomitante de condições mentais (American Psychiatric Association, 2013).

Estudos relatam que menos de 60% dos casos envolvendo ansiedade e/ou depressão são corretamente diagnosticados nos centros de primeiros socorros (HIRSCHFELD, 2001).

A popularização da Internet e o amplo uso de redes sociais promoveram oportunidades para implantar soluções computacionais para apoiar os estudos de transtornos mentais. Segundo Tuchlinkski (2018), um fator agravante para o tratamento de desordens mentais é a existência de preconceito em relação às pessoas que assumem publicamente estarem sofrendo de algum transtorno mental. Assim, muitas pessoas têm visto nas redes sociais uma alternativa para dar voz a suas dificuldades e encontrar apoio, uma vez que a identidade pode ser mantida anônima nesses serviços. Isso resulta em um volume crescente de dados de alto valor que podem ser explorados para reconhecer automaticamente transtornos mentais e seus critérios de diagnóstico, bem como a descoberta de interações entre esses distúrbios. Vários trabalhos contribuíram para a caracterização de transtornos mentais a partir de textos e interações disponíveis nas mídias sociais. Uma revisão sistemática (WONGKOBLAP; VADILLO; CURCIN, 2017) revela que os trabalhos relacionados se concentram na identificação automática de distúrbios específicos usando técnicas de aprendizado supervisionado.

A classificação automática de transtornos depressivos é o foco da maioria dos trabalhos. Uma revisão sistemática sobre estudos focados em depressão é apresentada em Giuntini et al. (2020). Esses estudos exploram tanto a classificação de usuários, quanto de *posts* usando uma abordagem de aprendizado supervisionada. Boa parte destes estudos aplicaram aprendizado a atributos (*features*) textuais, sociais e de sentimentos extraídos de dados, geralmente usando extensa engenharia de *features* (e. g. (CHOUDHURY et al., 2014; PARK et al., 2015; CHOUDHURY et al., 2017; CACHEDA et al., 2019)). Mais recentemente, os benefícios das técnicas de aprendizado profundo foram explorados para a classificação de depressão (YATES; COHAN; GOHARIAN, 2017; MANN; PAES; MATSUSHIMA, 2020) e ansiedade (SHEN; RUDZICZ, 2017). O aprendizado profundo tem o benefício de incluir a extração de representações de dados de entrada como parte do processo de aprendizado.

Poucos trabalhos abordam a classificação dos transtornos de ansiedade. Dos que o fazem, todos focam a classificação de *posts*. Embora alguns trabalhos abordem a classificação de múltiplas condições mentais, incluindo ansiedade e depressão (COPPERSMITH et al., 2015; GKOTSIS et al., 2017; BAGROY; KUMARAGURU; CHOUDHURY, 2017; IVE et al., 2018), raros trabalhos abordam a comorbidade entre essas desordens. Dos trabalhos que exploram a comorbidade, todos focam na classificação de usuários. Da-

dos de usuários auto-diagnosticados extraídos da rede social Reddit[2] para nove condições mentais, incluindo ansiedade, depressão e sua condição de comorbidade são explorados em (COHAN et al., 2018) sob a proposta de um modelo multirrótulo para classificação de usuários. Os autores exploraram as técnicas de aprendizado raso (e.g. FastText) e profundo (e.g. Convolutional Neural Networks (CNN)), sendo os resultados alcançados insatisfatórios em todas as abordagens. Outro estudo (BENTON; MITCHELL; HOVY, 2017), explorou o potencial do Multilayer Perceptron (MLP) como uma rede profunda para a classificação de usuários auto-diagnosticados da rede social Twitter[3], que apresentou bom desempenho para a identificação dos transtornos específicos, mas não abordou sua comorbidade.

Todos os trabalhos supracitados fornecem importantes contribuições para auxiliar na triagem de pacientes, bem como na identificação de padrões que permitam confirmar ou elucidar comportamentos já identificados em estudos realizados na área de saúde mental, mas que frequentemente envolvem um baixo número de participantes. Ainda, esses trabalhos confirmam a superioridade do aprendizado profundo na classificação de transtornos mentais a partir de dados de rede sociais. Contudo, os trabalhos realizados até o momento apresentam resultados insatisfatórios com relação ao modelo de classificação (COHAN et al., 2018), ou não exploram as relações de comorbidade entre as desordens em termos de padrões observados (BENTON; MITCHELL; HOVY, 2017). Especificamente, considerando os transtornos de ansiedade, não foram encontrados trabalhos que aprofundem a investigação das relações dessa desordem com a depressão em sua condição de comorbidade. O presente trabalho visa reduzir lacunas existentes no estudo de padrões para os transtornos de ansiedade e sua comorbidade com a depressão, com uma solução de classificação baseada em aprendizado profundo e comitês de classificadores.

Este trabalho tem como principal objetivo propor e avaliar modelos desenvolvidos para a identificação automática de depressão, ansiedade e sua comorbidade, a partir de um conjunto de dados de usuários diagnosticados extraído da rede social Reddit (COHAN et al., 2018). Para este fim, propõe-se *DAC Stacking*, um modelo que emprega a abordagem *ensemble* (comitê de classificadores) (MURPHY, 2012) para superar as dificuldades de lidar com um problema de classificação de várias classes e vários rótulos envolvido no cenário de comorbidade, onde os padrões de distinção são menos claros. No nível inferior, *DAC Stacking* é composto de classificadores binários (único rótulo) que visam distinguir entre usuários saudáveis (controle) e diagnosticados com uma das condições alvo espe-

[2]https://www.reddit.com/
[3]https://twitter.com/

cíficas. No nível superior, essas previsões individuais são consolidadas usando uma rede neural densa, que gerencia o problema de multirrótulo para a atribuição de rótulos de controle ou diagnosticados. Para atingir esse objetivo, diferentes alternativas de arquiteturas *ensemble* foram testadas para o *DAC Stacking*, explorando:

- *a função dos classificadores fracos*: foram experimentadas variações quanto ao uso de classificadores focados em distinguir usuários de controle daqueles com alguma condição condição-alvo (ansiedade, depressão e comorbidade), bem como classificadores diferenciadores de condições isoladas e sua comorbidade;

- *arquiteturas de aprendizado profundo*: foram testadas as redes do tipo CNN, Long-Short Term Memory (LSTM) e híbridas (formadas pela combinação dessas arquiteturas) para verificar se as distintas premissas de aprendizado contribuem de forma mais efetiva ao problema, quer de forma isolada ou combinada;

- *word embeddings*: foi avaliado tanto do uso de *word embeddings* pré-treinados em grandes quantidades de dados genéricos (e.g. Google News), quanto à contribuição de *embeddings* de domínio extraídos do próprio conjunto de dados Reddit, considerando o conteúdo de usuários diagnosticados com múltiplos transtornos.

Além da proposta do modelo de classificação, esse trabalho também busca fornecer *insights* que auxiliem na caracterização dos transtornos em isolados, e de sua comorbidade. Para isto, foram analisadas as relações entre as *features* influentes na classificação de cada condição mental, e os respectivos sintomas de acordo com manual de psicologia Diagnostic and Statistical Manual of Mental Disorders V (DSM-5) (American Psychiatric Association, 2013), que descreve distúrbios e seus sintomas. Para identificar *features* influentes na classificação, adotou-se o método SHapley Additive exPlanations (SHAP) (LUNDBERG; LEE, 2017), que explica a previsão de uma determinada instância de acordo com a teoria de jogos de coalizão. SHAP permite a interpretação global de características influentes pelas agregações de valores Shapley.

Os experimentos realizados tiveram objetivo de responder às seguintes questões de pesquisa:

- QP1: Uma solução do tipo *stacking ensemble* é efetiva para o problema de classificação multirrótulo envolvendo os transtornos de ansiedade, depressão e sua comorbidade?

- QP2: As distintas premissas de aprendizado, subjacentes às arquiteturas de aprendizado profundo exploradas (padrões locais, sequenciais), contribuem de forma isolada ou combinada à melhoria de desempenho da classificação em uma abordagem *stacking ensemble*?

- QP3: Qual topologia, em termos de função para os classificadores fracos e tipos de arquiteturas de aprendizado profundo, tem melhor desempenho nesta tarefa de classificação?

- QP4: Existe diferença de desempenho para esta tarefa de classificação advindo do uso de *word embeddings* pré-treinados genéricos e/ou de domínio?

- QP5: As *features* relevantes para a classificação de cada condição alvo representam características dos sintomas típicos de cada transtorno mental?

- QP6: Como a solução *DAC Stacking* proposta se compara a soluções estado da arte no tratamento de problemas na área de Processamento de Linguagem Natural (PLN)?

Todos os classificadores fracos binários utilizados no *DAC Stacking* superam as soluções existentes para depressão, ansiedade e comorbidade (YATES; COHAN; GOHARIAN, 2017; COHAN et al., 2018), considerando conjuntos de dados formados a partir da rede social Reddit. Os classificadores fracos baseados em CNN apresentaram o melhor desempenho, alcançando medida F de 0,79 para depressão, 0,78 para ansiedade e 0,78 para sua comorbidade. Todos os modelos *ensemble* também superaram as linhas de base. O melhor desempenho (*Hamming Loss* de 0,27 e *Exact Match Ratio* de 0,47) foi alcançado por uma topologia que combina os classificadores fracos LSTM, CNN e híbridos, e contém classificadores especializados com a função de diferenciadores entre as condições alvo. A interpretação qualitativa do *DAC Stacking* foi encorajadora, pois foi possível relacionar muitos recursos influentes aos sintomas descritos no manual de psicologia do DSM-5.

As principais contribuições desta pesquisa são:

- uma abordagem *stacking ensemble* que combina diferentes arquiteturas de aprendizado profundo para solucionar o problema de classificação multirrótulo envolvendo usuários saudáveis e diagnosticados com ansiedade, depressão ou ambos transtornos. Os experimentos realizados demonstram que a topologia que combina diferentes arquiteturas e funções para os classificadores binários apresenta o melhor

desempenho. Os resultados preliminares foram discutidos em (SOUZA; NOBRE; BECKER, 2020);

- um extenso conjunto de experimentos e análise comparativa de desempenho entre as arquiteturas de aprendizado profundo LSTM, CNN e modelos híbridos derivados da etapa de composição da topologia para os classificadores fracos do *DAC Stacking*. Esses experimentos revelam que a variação de *word embeddings*, de arquiteturas e de funções para os classificadores fracos contribuem para aumentar a variabilidade da solução *stacking*;

- *insights* para o estudo dos transtornos mentais alvo desse trabalho, através de uma análise sobre os padrões e relações encontradas entre as desordens. Para tanto, propôs-se um método de avaliação qualitativa baseado em *embeddings* e SHAP, que permitiu estabelecer uma relação entre os termos destacados como mais relevantes pelos classificadores fracos e os sintomas que identificam cada transtorno, segundo o manual DSM-5.

O restante deste trabalho está organizado como segue. O Capítulo 2 apresenta uma fundamentação teórica para a compreensão do trabalho proposto. O Capítulo 3 apresenta uma revisão dos trabalhos realizados na área. O Capítulo 4 apresenta o processo de formação e estatísticas sobre os conjuntos de dados usados neste trabalho. O Capítulo 5 apresenta estrutura do trabalho proposto. O Capítulo 6 apresenta resultados e avaliações qualitativas e quantitativas. O Capítulo 7 apresenta conclusões, limitações e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

O trabalho proposto visa identificar traços de ansiedade, depressão e sua comorbidade, em indivíduos a partir de sua comunicação em redes sociais. Para atingir esse objetivo, é necessário conhecer o processo de caracterização e diagnóstico dessas desordens mentais. Também são necessários conhecimentos sobre técnicas computacionais para solução de tarefas de processamento de linguagem natural. As seções deste capítulo são dedicadas a elucidar, brevemente, os conceitos necessários para compreensão do trabalho proposto.

2.1 Transtornos Mentais e Processo de Diagnóstico

Um transtorno mental é definido como uma síndrome caracterizada por perturbação clinicamente significativa na cognição, na regulação emocional ou no comportamento de um indivíduo, e que reflete uma disfunção nos processos psicológicos, biológicos ou de desenvolvimento subjacentes ao funcionamento mental (American Psychiatric Association, 2013). Segundo relatório da OMS (WHO, 2017), os transtornos depressivos e de ansiedade estão entre as doenças com maior impacto incapacitante na sociedade.

Os transtornos depressivos são distúrbios mentais caracterizados pela presença de humor triste, vazio ou irritável, acompanhado de alterações somáticas e cognitivas que afetam significativamente a capacidade do indivíduo de funcionar, prejudicando seu desempenho nas tarefas diárias e na vida social (American Psychiatric Association, 2013). Conforme critérios de duração, momento ou etiologia presumida, os transtornos depressivos são categorizados em oito tipos, com destaque para transtorno disruptivo de desregulação do humor, transtorno depressivo maior (incluindo episódio depressivo maior) e transtorno depressivo persistente (distimia) (HEALTH; EXCELLENCE, 2011).

Os transtornos de ansiedade incluem transtornos que compartilham características de medo e ansiedade excessivos em vários domínios (crianças, violência, profissão, etc.). O medo é definido como a resposta emocional à ameaça iminente real ou percebida, enquanto ansiedade é a antecipação de ameaça futura. Os ataques de pânico se destacam dentre os transtornos de ansiedade como um tipo particular de resposta ao medo (American Psychiatric Association, 2013). Além desse, o indivíduo pode experimentar outros sintomas físicos que incluem inquietação ou sensação de tensão; dificuldade de concentração; irritabilidade; tensão muscular e distúrbios do sono. Os transtornos de ansiedade

diferem entre si nos tipos de objetos ou situações que induzem medo, ansiedade ou comportamento de esquiva e na ideação cognitiva associada. São categorizados, segundo critérios específicos, em onze tipos. Os tipos mais frequentemente diagnosticados são o Transtorno de Ansiedade Generalizada, Transtorno de Pânico e Transtorno de Ansiedade Social (Fobia Social) (HEALTH; EXCELLENCE, 2011).

O diagnóstico de um transtorno mental pode seguir diferentes paradigmas, conforme o campo científico que o analisa. A psiquiatria segue o modelo médico, no qual o diagnóstico é resultado do processo empírico de observação, descrição e categorização de enfermidades que compartilham sinais e sintomas. Já para a análise do comportamento, a formulação de um diagnóstico passa pela compreensão dos comportamentos que são tidos como inadequados, e para tanto, as circunstâncias em que o comportamento se manifesta são de suma importância. Contudo, as diferentes formas de análise para os transtornos mentais não são conflitantes e, em alguns casos, podem ser complementares, desde que haja clareza sobre o tipo de informação obtido de cada análise (ARAUJO; NETO, 2014).

Independente do paradigma adotado para o diagnóstico, os profissionais de saúde mental contam com o manual de psicologia DSM-5 para apoiar suas decisões (SECAD, 2018). Em sua quinta edição, DSM-5 (American Psychiatric Association, 2013), fornece uma visão ampla sobre os transtornos mentais, baseada na contribuição de diferentes áreas da saúde, além de uma padronização para a linguagem clínica. Considerado por muitos especialistas como a principal referência no assunto (SECAD, 2018), esse manual contém um relato descritivo sobre os critérios que permitem identificar cada transtorno mental. Esses critérios descrevem não somente o conjunto de possíveis sintomas físicos e comportamentais, mas também a frequência e periodicidade com as quais devem se manifestar para que o indivíduo seja considerado diagnosticado com determinado transtorno mental. Além da análise do relato do paciente, os especialistas podem aplicar instrumentos de avaliação psicométricos de apoio ao diagnóstico, os quais sintetizam os critérios em um conjuntos de questões, que auxiliam na identificação dos sintomas, bem como sua intensidade e frequência de ocorrência (CICCHETTI, 1994).

Alguns sintomas e comportamentos são comuns a mais de uma desordem. Para esses casos, o contexto e, até mesmo a frequência e periodicidade dos sintomas, podem ajudar a definir o quadro clínico do paciente. Com relação à importância do contexto associado aos sintomas é possível verificar, por exemplo, que tanto indivíduos com transtorno depressivo maior quanto transtorno de ansiedade social, podem manifestar preocupação em serem avaliados negativamente por outras pessoas. Para os indivíduos depressivos,

esse sintoma está relacionado ao sentimento de considerarem-se maus ou não merecedores de afeto (American Psychiatric Association, 2013). Para os indivíduos ansiosos, no entanto, essa preocupação é justificado pelo medo de apresentar certos comportamentos sociais ou sintomas físicos em público. Com relação à frequência e periodicidade dos sintomas verifica-se, por exemplo, que a decisão entre o diagnóstico para o transtorno de depressão persistente (Distimia) e o transtorno de depressão maior, pode ser tomada com base se o paciente experienciou ou não um episódio de depressão maior nos últimos dois anos. Para os casos negativos, o paciente pode ser diagnosticado como distímico (American Psychiatric Association, 2013).

Para os casos em que não é possível fazer distinção, as desordens são identificadas juntamente com o termo comorbidade, indicando sua co-ocorrência. Segundo a American Psychiatric Association (2013), a ansiedade é um dos transtornos que apresentam maior comorbidade com outros transtornos mentais, entre eles os transtornos depressivos. Alguns estudos apresentam evidências da relação próxima entre ansiedade e depressão (HIRSCHFELD, 2001). A taxa de comorbidade de ansiedade e depressão é alta, sendo que 85% dos pacientes com depressão também apresentam sintomas significativos de ansiedade, enquanto a depressão ocorre em mais de 90% dos pacientes que sofrem de transtornos de ansiedade (TILLER, 2013). Essa comorbilidade acentua o quadro clínico do indivíduo deprimido, levando a um maior risco de suicídio, pior funcionamento social e resistência ao tratamento (HIRSCHFELD, 2001).

Neste trabalho abordaremos a ansiedade, depressão e a condição de comorbidade entre essas desordens a partir de conteúdo extraído da rede social Reddit. Para a análise dos padrões identificados e suas relações, adotamos o manual DSM-5, por ser considerado a principal referência no assunto.

2.2 Processamento de Linguagem Natural e o Aprendizado de Máquina

A evolução das técnicas de aprendizado de máquina vem propiciando importantes avanços na área de PLN. Em (DENG; LIU, 2018), PLN é caracterizado como um campo de pesquisa interdisciplinar dedicado ao estudo do uso de computadores para processar e/ou compreender a linguagem humana (natural), combinando conhecimentos nas áreas de linguística computacional, ciência da computação, ciência cognitiva e inteligência artificial. Aplicações típicas dessa área incluem, entre outras, o reconhecimento e compreensão da linguagem falada, análise lexical, análise de sentimentos e computação social.

Entre os principais avanços em aprendizado de máquina que beneficiaram a área de PLN, destaca-se a aplicação de aprendizado profundo para o desenvolvimento de (1) representações distribuídas de entidades linguísticas via *embeddings*, o que propiciou associação de contexto para cada palavra (generalização semântica); (2) redes hierárquicas eficazes para representar diferentes níveis linguísticos e sequencias profundas de longo alcance da linguagem natural; e (3) métodos *end-to-end* para resolver múltiplas tarefas em PLN.

O aprendizado de máquina é definido por Murphy (2012) como uma subárea da inteligência artificial dedicada ao desenvolvimento de métodos ou técnicas computacionais capazes de detectar automaticamente padrões em dados e, a partir do reconhecimento desses padrões, predizer dados futuros ou realizar outros tipos de tomada de decisão sob cenário de incerteza. Entre os tipos de aprendizado de máquina, destaca-se o supervisionado, no qual o algoritmo aprende com base em exemplos, devidamente rotulados e fornecidos durante o treinamento do modelo. Essa categoria de algoritmos é a mais empregada para solução de tarefas de classificação na área de PLN (MURPHY, 2012). Considerando a classificação de transtornos mentais a partir de conteúdo extraído de redes sociais, verifica-se tanto a exploração de técnicas de aprendizado supervisionado raso, quanto profundo. As seções seguintes resumem as técnicas mais usadas para essa categoria de problemas PLN.

2.2.1 Aprendizado Raso e Engenharia de *Features*

Existe um vasto conjunto de técnicas de aprendizado raso, sendo os métodos baseados em otimização Support Vector Machine (SVM), procura (regressões e árvores de decisão) e probabilísticos (métodos baseados no teorema de Bayes), bastante usados em tarefas de PLN envolvendo a classificação de texto. Os principais algoritmos para cada categoria de métodos são (MURPHY, 2012):

- *SVM*: método de *kernel* que busca encontrar um hiperplano que melhor divida os pontos de dados, segundo as classes alvo. Para tanto, o algoritmo realiza um mapeamento dos dados de treinamento para uma nova representação de alta dimensão e calcula a máxima distância entre os hiperplanos gerados e os pontos de dados mais próximos para cada classe;

- *Árvores de Decisão*: modelo hierárquico capaz de guiar a tomada de decisão sobre a qual classe pertence uma determinada instância, resultando um caminho único

do nó raiz à folha (classe alvo). O modelo de árvore é obtido a partir dos dados de treinamento usando uma estratégia de divisão e conquista, aplicada hierarquicamente. Nessa categoria de algoritmos, destacam-se na atualidade o *Random Forest* e *Gradient Boosting Machines*;

- *Regressão*: método preditivo que analisa a relação entre variáveis dependentes (alvo) e independentes (preditoras). Diferentes algoritmos são usados conforme o tipo da variável alvo e/ou a relação entre variáveis (linear ou não linear). Dentre eles, destacam-se Regressão Linear, Regressão Logística e Regressões Bayesianas.

De modo geral, as técnicas de aprendizado raso envolvem a conversão dos dados de entrada em um ou mais espaços de representação sucessivos. No entanto, esses espaço de hipótese gerados não são suficientemente ricos para permitir o aprendizado de representações mais refinadas, as quais são necessárias para solução de problemas complexos. Para superar essa dificuldade e obter bons resultados com aprendizado raso, essas representações passaram a ser definidas manualmente em um processo denominado de engenharia de *features* (DENG; LIU, 2018). Isso abrange ações de pré-processamento, inclusão de *features* definidas partir de conhecimentos específicos sobre o domínio do problema e, em muitos casos, aplicação de técnicas de redução de dimensionalidade para melhorar o desempenho da classificação (MURPHY, 2012).

As ações de pre-processamento compreendem:

- *Uniformização*: processo para determinar um formato único para os termos (e.g. conversões em letra minúscula, uso de hifenização para nomes compostos);

- *Limpeza de dados*: envolvem ações como remoção de pontuação, caracteres especiais e *stopwords*, que são palavras de pouco significado para representar um documento (e.g. artigos, preposições e conjunções);

- *Stemming e Lematização*: o método *stemming* visa reduzir as formas variantes das palavras por remoção de afixos e sufixos, de modo que apenas o radical, que contém o significado do termo, é mantido. Diferentemente, a lematização é um processo algorítmico realizado para determinar o lema de uma palavra com base no significado e contexto em que é empregada. Esse processo resulta em grupos de termos que podem ser sintaticamente diferentes, porém usadas com o mesmo significado;

- *Vetorização de Dados*: processo para conversão do conteúdo textual em numérico. Consiste em aplicar um método de tokenização para segmentar o texto em unidades

menores, denominadas *tokens*, as quais podem ser palavras, caracteres ou n-gramas. Em seguida, para cada *token* gerado atribui-se um vetor numérico, conforme o método de associação selecionado; Os métodos mais comuns usados na etapa vetorização de dados incluem (1) termos mais frequentes no texto, sem preservar a ordem que são apresentados em cada sentença (*bag of words*); (2) termos mais frequentes entre os documentos (Document Frequence - DF); (3) termos raramente encontrados entre documentos (Inverse Document Frequence - IDF); (4) combinação desses métodos (TF-IDF). Outro método robusto para associar um vetor a uma palavra é o uso de vetores de palavras densas, também chamados de *word embeddings*, os quais são vistos em detalhes na Seção 2.3.

A engenharia de *features* também pode incluir a geração de *features* específicas ao domínio, que variam conforme o tipo de problema. Considerando a análise de transtornos mentais a partir de conteúdo de redes sociais, por exemplo, podem ser geradas estatísticas a partir de recursos próprios da rede social em uso (e.g. total de *likes*, *posts*, etc), combinadas ao emprego de ferramentas para análise de diferentes estruturas a partir de texto. Ferramentas bastante utilizadas para a análise textual são os dicionários de sinônimos (e.g WordNet (FELLBAUM, 1998)) ou de sentimento, tais como o Linguistic Inquiry and Word Count (LIWC) (GONZALES; HANCOCK; PENNEBAKER, 2010), o Affective Norms for English Words (ANEW) (BRADLEY et al., 1999) e o Valence Aware Dictionary and sEntiment Reasoner (VADER) (HUTTO; GILBERT, 2014). Em comum, esses dicionários classificam as palavras segundo os critérios de composição da linguagem (e.g. classes gramaticais) e/ou processos psicológicos que incluem análise emocional, cognitiva e social. Essa classificação é usada como auxiliar para a identificação dos possíveis comportamentos, contextos e/ou sentimentos associados ao uso de determinadas palavras. Estas *features* específicas ao domínio também são incluídas no vetor de representação da instância de dados, através de pesos binários ou escalares.

Por fim, o vetor resultante da definição do conjunto de *features* para a representação de cada instância de dados pode apresentar alta dimensionalidade, tornando a etapa de aprendizado muito custosa (ou mesmo impraticável), degradando o desempenho. Nesses casos, técnicas de redução de dimensionalidade são aplicadas para reduzir a quantidade de atributos analisados, projetando os dados em um subespaço de dimensão inferior que descreve a maior parte da variabilidade dos dados (MURPHY, 2012). Essas técnicas podem ser divididas em duas grandes abordagens (GAMA et al., 2011): agregação e seleção de *features*. Na primeira abordagem, a redução da dimensão dos dados é realizada pela

combinação dos atributos originais usando funções lineares e não lineares. Entre as técnicas baseadas nessa abordagem, destacam-se a (1) Principal Components Analysis (PCA) e (2) Linear Discriminant Analysis (LDA). Na seleção de *features*, diferentes técnicas podem ser empregadas para seleção automática de atributos. De modo geral, essas técnicas buscam um subconjunto ótimo de atributos de acordo com um dado critério usado para medir o poder preditivo do atributo. Entre as diferentes técnicas, destacam-se as abordagens baseadas no critério de ganho de informação por entropia, correlação estatística e algoritmos genéticos.

2.2.2 Aprendizado Profundo

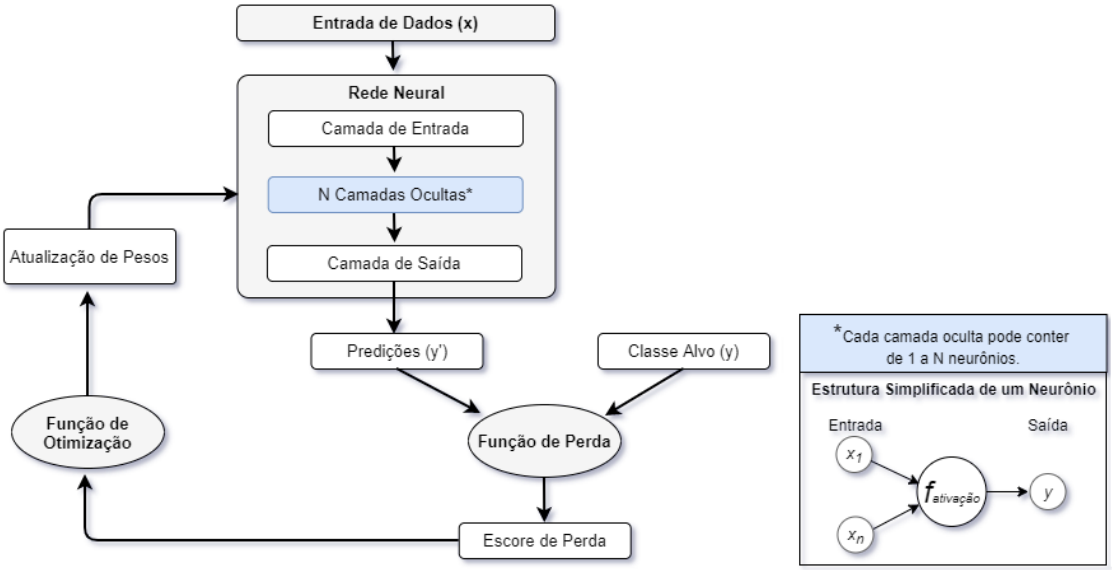
No aprendizado profundo, a extração e a transformação dos dados de entrada em *features* é parte integrante do processo de aprendizado, desenvolvido a partir de uma cascata de várias camadas de unidades de processamento não lineares (GOLDBERG, 2017). Enquanto as camadas inferiores aprendem as *features* diretamente sobre os dados de entrada, as camadas superiores aprendem as *features* a partir de derivações mais complexas. Essa estrutura em camadas produz uma representação de recursos poderosa que cresce hierarquicamente em complexidade (MURPHY, 2012).

Em relação às representações em camadas hierárquicas, geralmente são aprendidas por meio de modelos de aprendizado profundo denominados redes neurais, estruturados segundo empilhamento de camadas. De modo geral, uma rede neural executa um mapeamento dos dados de entrada para o dado alvo, por exposição de inúmeros exemplos de dados. Cada camada realiza transformações simples, que ao longo de muitos estágios são capazes de identificar conexões mais complexas para as representações de características extraídas dos dados de entrada. A Figura 2.1(a) ilustra a relação entre os principais componentes presentes em uma rede neural, destacando sua unidade fundamental (neurônio) na Figura 2.1 (b). Entre esses componentes, destacam-se os seguintes elementos (GOLDBERG, 2017):

- *Camada de entrada*: define o formato de entrada dos dados pré-processados (vetores numéricos) conforme a arquitetura das camadas ocultas, as quais podem ser beneficiadas por subdivisões nos dados de entrada;

- *Camadas ocultas*: responsáveis por aplicar transformações nos dados de entrada, as quais são parametrizadas por um conjunto de pesos. Nesse contexto, aprender

Figura 2.1: Relação entre os principais componentes de uma rede neural.



(a) Principais Componentes (b) Neurônio

Fonte: Adaptado de Chollet (2017)

significa encontrar um conjunto de valores para os pesos de todas as camadas de uma rede, de modo que a rede mapeie corretamente as entradas de exemplo para seus alvos associados. Diferentes hiperparâmetros podem ser definidos para as camadas ocultas, conforme o tipo de rede neural. Entretanto, toda a camada oculta deve conter uma definição do número total de unidades (neurônios) e da função de ativação, responsável por transformar os dados de entrada em saída;

- *Função de Ativação*: implementa transformações não lineares sobre os dados de entrada de uma camada, resultando na saída de novos dados. As funções não lineares possibilitam a cada camada gerar espaços de hipóteses mais ricos, necessários para problemas que exigem representações mais complexas dos dados. Essas funções são definidas tanto nas camadas ocultas da rede, quanto na camada de saída. Entre as funções de ativação mais usadas para camadas ocultas, destacam-se (1) Unidade Linear Retificada (ReLU), a qual mapeia valores negativos para zero; e (2) Tangente Hiperbólica (Tanh) que atua na conversão dos dados de entrada para um intervalo de valores entre -1 e 1. Para a camada de saída são muito usadas as funções (1) *Softmax*, que gera uma distribuição de probabilidades que varia de 0 a 1, representando a dependência de ocorrência entre todas as classes possíveis; e (2) *Sigmoid*, usada para problemas em que a probabilidade de ocorrência de uma classe é independente das demais;

- *Função de Perda*: também conhecida como função de custo *J*, é responsável por

medir a perda do modelo (gradiente de perda), estimada em termos da diferença entre o valor predito e o valor real da classe para cada amostra de treinamento. Entre os tipos de função de perda, destacam-se as funções *crossentropy*, as quais medem o desempenho para modelos de classificação cuja saída é um valor de probabilidade entre 0 e 1. Valores maiores indicam maior diferença entre os valores predito e real;

- *Função de Otimização*: especifica o modo como o gradiente de perda será usado para atualizar os pesos da rede, de modo a minimizar a perda do modelo. Os principais algoritmos para retro-propagação dos gradientes são *Adam*, *Adadelta*, *Adagrad*, *Adamax* e *Nadam*. Entre os parâmetros ajustáveis para essas funções, destaca-se a taxa de aprendizado, cujo valor determina o tamanho do passo para a atualização dos parâmetros na direção oposta ao gradiente da função de perda. Uma taxa de aprendizado adequada pode reduzir o tempo de convergência do modelo;

- *Regularização*: abordagem empregada para evitar o (*overfitting*) das redes neurais durante a etapa de treinamento. Consiste no emprego de técnicas para reduzir a quantidade de parâmetros ou resultados propagados pela rede durante o aprendizado do modelo. O termo de regularização considera os valores dos parâmetros e pontua sua complexidade, visando encontrar parâmetros que apresentem baixa perda e complexidade. O parâmetro regulador permite ponderar a escolha entre modelos simples ou de baixa perda. Entre as técnicas de regularização destaca-se o *Dropout*. Esse método atua no descarte dos resultados produzidos pelos neurônios, ou por camadas específicas da rede, conforme percentual de descarte definido. Nesse método, as posições para descarte na matriz de resultados são selecionadas de modo aleatório;

- *Camadas de Saída*: definida conforme o tipo de tarefa que o modelo de rede neural visa solucionar. Independente da tarefa, essa camada deve ser configurada quanto ao total de unidades e as funções de ativação e perda. Considerando as tarefas de classificação, as unidades são definidas segundo o total de classes possíveis. Quanto às funções de ativação e perda, sua escolha está relacionada (1) ao número de classes definidas para o problema (binário ou multi-classe); e (2) à definição de quantas classes podem ser previstas para uma instância (rótulo único ou multirrótulo). Assim, para problemas de classificação multirrótulo ou binários, a função de ativação mais indicada é a *sigmoid* e a função perda mais indicada é *binary crossentropy* (CHOLLET, 2017).

2.2.3 Redes Neurais Recorrentes

Um tipo de rede neural que vem sendo aplicado com sucesso para tarefas de classificação de texto são as Recurrent Neural Network (RNN). Essas redes diferem das demais por introduzirem um mecanismo de memória recente via inclusão de conexões de *feedback* (MURPHY, 2012). O modelo LSTM (HOCHREITER; SCHMIDHUBER, 1997) é uma RNN que expande esse mecanismo de memória para tratar com sequencias mais longas de dados, garantindo que todas as informações necessárias fiquem disponíveis para cada estágio da rede, prevenindo a perda de sinal, efeito conhecido como *vanishing gradient*. O problema de *vanishing gradient* ocorre quando o gradiente é reduzido a valores muitos pequenos a medida que são retro-propagados para a atualização dos pesos da rede, o que resulta na perda de aprendizado para camadas anteriores.

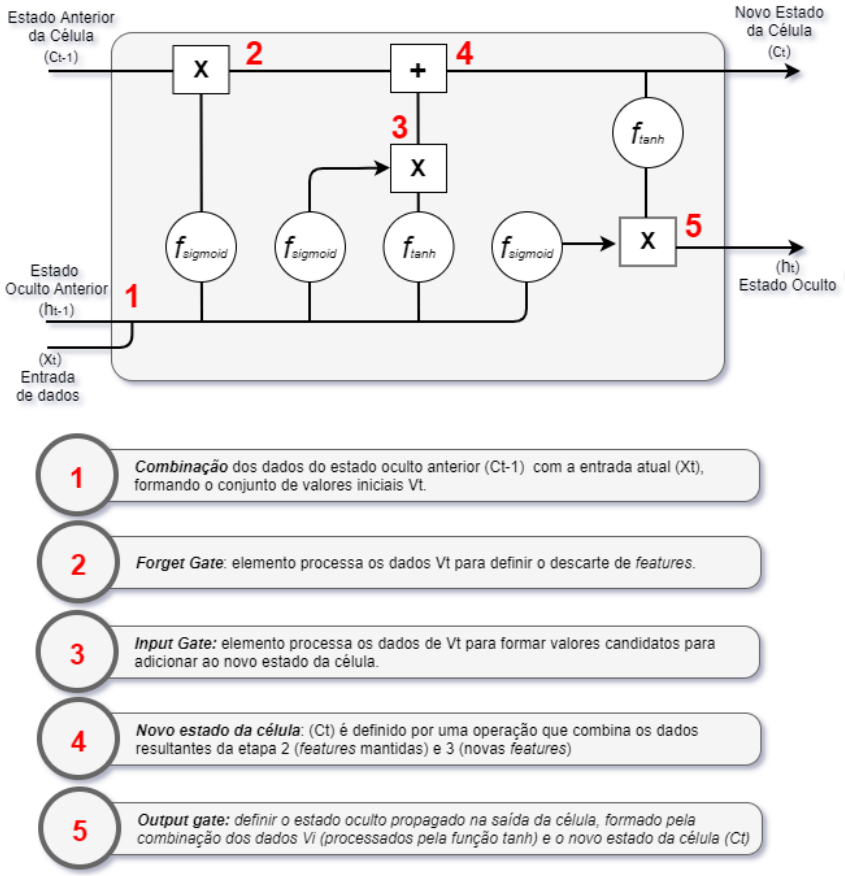
A Figura 2.2 ilustra a estrutura de uma célula LSTM juntamente com a sequencia de operações realizadas durante a etapa de treinamento. Considerando o processamento de textos, o processo de aprendizado das redes LSTM tenta reproduzir, de modo simplificado, o mecanismo de leitura humana. Nele, um texto é lido palavra a palavra de modo sequencial. De modo semelhante, a rede LSTM processa uma sequencia de dados ao longo de várias iterações. A cada iteração, o estado da célula LSTM é atualizado segundo operações realizadas por elementos denominados *gates*. Essas operações incluem combinação do estado atual e dados de entrada, seleção de *features* não relevantes para descarte e de features novas, identificadas a partir da entrada. Ao final da sequencia de etapas, a célula LSTM propaga seu estado de memória atualizado juntamente com o estado oculto da camada.

As RNNs podem ser formadas por várias camadas completamente conectadas de células LSTM. A formação de modelos a partir da arquitetura LSTM envolve a definição de um conjunto de parâmetros específicos (hiperparâmetros) que visam ajustar esse mecanismo de memória e iterações a sequencia de dados fornecida na entrada da rede. Entre os diferentes hiperparâmetros disponíveis, destacam-se os explorados por esse trabalho (GANEGEDARA, 2018):

- *Dropout*: ou abandono, permite definir a taxa de descarte das *features* durante a etapa de transformação linear das entradas, com a finalidade de combater o superajuste (*overfitting*) durante o treinamento da rede;

- *Recurrent Dropout*: determina a taxa de descarte das *features* aplicando a mesma

Figura 2.2: Estrutura e Etapas de Operação da Célula LSTM



Fonte: Adaptado de Chollet (2017)

técnica de regulação explanada anteriormente, porém considerando a operação de transformação linear entre os estados recorrentes em uma camada LSTM;

- *Return Sequence*: refere-se ao retorno do estado oculto gerado a cada passo de treinamento (*timestep*), considerando um mesmo lote (*batch*) de dados. A ativação desse parâmetro determina se a sequencia completa de estados ocultos sucessivos, produzidos por cada *timestep*, deve ou não ser retornada pela célula LSTM. Quando desabilitado, retorna somente o estado oculto do último *timestep*, o qual captura uma representação mais abstrata da sequencia de entrada. Esse parâmetro normalmente é ativado quando tem-se mais de uma camada LSTM na rede, com o intuito de propagar para a camada seguinte, o conjunto de informações produzido pela extração de dados em cada *timestep*;

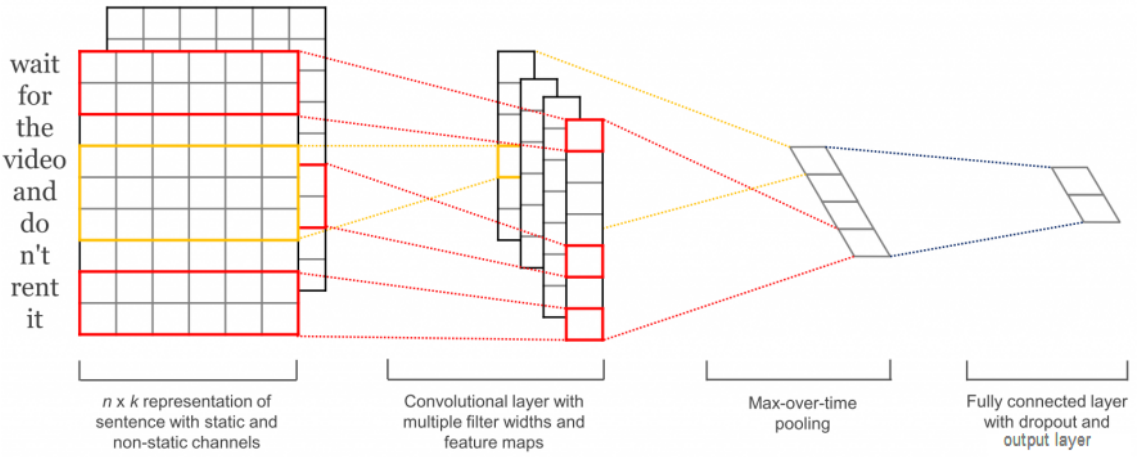
- *Stateful*: determina se o estado da célula LSTM, ao final do lote de treinamento, deve ser mantido como estado inicial para o próximo lote de dados. O padrão da rede LSTM é manter esse parâmetro desabilitado, forçando a inicialização dos pesos da rede a cada novo lote de dados.

Outros tipos de RNN aplicados em tarefas de classificação de texto, incluem os modelos (1) Gated Recurrent Unit (GRU) (CHO et al., 2014), uma versão simplificada de LSTM com menor custo computacional; e (2) Embeddings from Language Models (ELMo) (PETERS et al., 2018b), cujo modelo é formada por sucessivas camadas LSTM Bi-direcionais.

2.2.4 Redes Neurais Convolucionais

As redes neurais *feedforward*, também conhecida como Multilayer Perceptron (MLP), são compostas por uma série de camadas empilhadas, cada uma formada por um modelo de regressão logística, seguidos por uma camada final formada por um modelo de regressão logística ou linear, dependendo se a solução do problema é para uma tarefa classificação ou regressão. Uma forma particular de rede MLP é a Convolutional Neural Network (CNN) (MURPHY, 2012). Diferente das RNNs, esse tipo de rede não contém mecanismo de memória. Assim, cada entrada de dados é processada de forma independente, sem nenhum estado mantido entre as entradas. Desse modo, para processar uma sequência ou série temporal de pontos de dados, é necessário mostrar a sequencia completa para a instância, transformando-a em um único ponto de dados. Conforme a dimensão dos dados de entrada, a rede CNNs é identificada como 1D (sequencias), 2D (imagens) e 3D (volumes) (CHOLLET, 2017). A Figura 2.3 ilustra a estrutura de uma rede CNN 1D, aplicada a sequências de dados textuais.

Figura 2.3: Estrutura das Redes Neurais Convolucionais 1D



Fonte: Britz (2016)

De modo geral, o processo de aprendizado das redes CNN é baseado no emprego

de sucessivas operações de convolução a partir dos dados de entrada. Essas operações são realizadas pela camada convolucional, a qual é definida principalmente pelos parâmetros:

- *kernel size*: define o tamanho da janela de convolução, ou seja, a região de dados;
- *strides*: determina o passo de convolução, isto é, a distância entre duas janelas para análise de dados;
- *filters*: o total de filtros usados define a dimensionalidade do espaço de saída na convolução.

A camada convolucional é responsável por aplicar os filtros sobre cada região gerada pelo deslocamento da janela de convolução. Nessa estrutura, os filtros atuam como um detector de padrões locais, buscando por janelas que contenham *features* específicas. Considerando dados textuais, as janelas são fragmentos de texto de k termos. Ao final desse processo, diferentes mapas de *features* são derivados. O aprendizado de padrões não considera a localização dos fragmentos. Desse modo, padrões aprendidos a partir de um fragmento podem ser reconhecidos em outras sequências de texto.

Cada camada convolucional é seguida por uma camada de *pooling*, a qual é responsável por reduzir a dimensionalidade dos mapas de *features* gerados, através da eliminação de regiões sobrepostas. A camada de *pooling* é nomeada conforme o critério de seleção de *features*. As opções mais usadas são *Max Pooling* e *Global Average Pooling*, as quais consideram o valor máximo e médio, respectivamente, para a seleção de *feature* em cada região.

Neste trabalho, as redes CNN 1D são exploradas para a formação de modelos de classificação de usuários a partir do conteúdo de suas postagens na rede social Reddit. Para ajustar a rede CNN a esse problema de classificação, foram explorados os parâmetros específicos *kernel size* e *filters*. O parâmetro *strides* foi mantido com a configuração padrão [1].

2.3 *Word Embeddings*

As incorporações de palavras ou *word embeddings* são uma abordagem robusta que permite associar palavras com vetores de ponto flutuante de baixa dimensão, os quais são aprendidos a partir dos dados. Essas representações distribuídas são apreendidas

[1] Disponível em: https://keras.io/api/layers/convolution_layers/convolution1d/

com base no uso das palavras, o que permite que palavras usadas de maneira semelhante resultem em representações semelhantes, capturando mais naturalmente seu significado (DENG; LIU, 2018).

Os *word embeddings* podem ser obtidos durante o treinamento do modelo para a tarefa principal (e.g. classificação de documentos), ou pré-treinados, como uma tarefa separada. Na primeira opção, os vetores de palavras são iniciados com valores aleatórios e ajustados durante o treinamento, junto com os demais pesos da rede neural. Já os *embeddings* pré-treinados, são previamente computados usando uma tarefa de aprendizado de máquina diferente da tarefa principal.

Os modelos para aprendizado de *embeddings* de palavras compreendem duas categorias (1) métodos baseados em janela de contexto local, os quais pressupõem que as palavras que compartilham palavras semelhantes ao redor são semanticamente próximas, sendo um dos mais usados Word2Vec (MIKOLOV et al., 2013); e (2) métodos baseados em fatoração de matriz global, os quais exploram estatísticas globais de um corpus (e.g. co-ocorrência de palavras), onde destaca-se a técnica GloVe (PENNINGTON; SOCHER; MANNING, 2014a).

A técnica Word2Vec (MIKOLOV et al., 2013) propõe o uso de redes neurais *feedforward* para o aprendizado das representações distribuídas de palavras, segundo duas abordagens: Continuous Bag-of-Words (CBOW) e Skip-gram. No algoritmo CBOW, o objetivo de treinamento do modelo é combinar as representações das palavras circundantes, cujo total é definido por uma janela de tamanho fixo, para prever a palavra alvo. Já o algoritmo de Skip-gram usa a palavra central para predizer as palavras circundantes. Ambos algoritmos são implementados pela biblioteca *Gensim*[2] usada para o treinamento de *embeddings* pré-treinados, a partir de grandes corpora, tais como o Google News[3]. A Tabela 2.1 apresenta os principais hiperparâmetros disponíveis para configuração do modelo Word2Vec.

Tabela 2.1: Word2Vec: Principais hiperparâmetros do modelo.

Hiperparâmetro	**Descrição**
Embedding Size	Dimensões da incorporação, o comprimento do vetor denso para representar cada token (palavra).
Window	A distância máxima entre uma palavra-alvo e palavras circundantes (contexto).
Minimum count	A contagem mínima de palavras a considerar durante o treinamento. Palavras extremamente raras geralmente não tem relevância em um grande conjunto de dados.
Iteration	Número de passagens que o algoritmo faz através do conjunto de dados.
Workers	Número de *threads* a serem usados durante o treinamento
Algorithm	Skip-gram ou CBOW

[2]Biblioteca disponível em https://radimrehurek.com/gensim/models/word2vec.html
[3]Disponível em https://code.google.com/archive/p/word2vec/

Tabela 2.2: Glove: Principais hiperparâmetros do modelo.

Hiperparâmetro	**Descrição**
Window	Distância considerada entre duas palavras para encontrar alguma relação entre elas.
No of components	Dimensão do vetor de sáida gerado pelo algoritmo Glove.
Learning Rate	Taxa de aprendizado para o decaimento do gradiente descendente. Quanto menor, maior o tempo de treinamento.
Epochs	Número de passagens que o algoritmo faz através do conjunto de dados.
No of Threads	Número de *threads* a serem usadas durante o treinamento.

O GloVe (PENNINGTON; SOCHER; MANNING, 2014b) é um algoritmo de aprendizagem não supervisionado para obter representações vetoriais de palavras. O treinamento é realizado usando estatísticas globais agregadas de co-ocorrência palavra-palavra de um corpus, e as representações resultantes mostram subestruturas lineares interessantes do espaço vetorial de palavras. O algoritmo GloVe é implementado pela biblioteca *glove*[4] e foi usado para gerar *embeddings* pré-treinados a partir de grandes quantidades de dados, tais como GloVe 6B[5] e GloVe Twitter[5]. A Tabela 2.2 destaca os principais hiperparâmetros para treinamento desse algoritmo.

A inclusão de uma camada de *embeddings* à rede neural profunda implica a definição de seu uso quanto à abordagem de aprendizado. Entre as técnicas existentes, destacam-se o aprendizado *estático* e *não estático*. A diferença entre essas abordagens reside em manter os *embeddings* conforme seu estado inicial (estático) ou permitir que sejam ajustados durante o processo de treinamento do modelo (não-estático).

Neste trabalho, *embeddings* pré-treinados foram usados na formação dos classificadores dedicados a identificação das condições de ansiedade, depressão e comorbidade. O conjunto de *embeddings* pré-treinados experimentados inclui os de propósito geral (Glove 6B, Twitter e Google News) e os de domínio, os quais foram gerados a partir do corpus, segundo as técnicas Word2Vec (Skip-gram e CBOW) e GloVe. Cada *embedding* foi utilizando tanto com abordagem de aprendizado estático, quanto não estático.

2.4 BERT

O Bidirecional Encoder Representations from Transformers (BERT) é uma técnica de aprendizado profundo baseada em *Transformers* (VASWANI et al., 2017) para criar modelos de representação de linguagens, sendo treinado em quantidades massivas de dados. BERT é o atual estado da arte para muitas aplicações em PLN, tais como classificação, pergunta e resposta, e identificação de entradas nomeadas. A inovação chave

[4]Biblioteca disponível em https://pypi.org/project/glove/
[5]Disponível em https://nlp.stanford.edu/projects/glove/

dessa técnica é a aplicação de uma abordagem de treinamento bidirecional para o *transformer*. Segundo Devlin et al. (2019), a abordagem BERT para o treinamento dos *transformers* resulta em um senso mais profundo do contexto e fluxo da linguagem do que as representações geradas por alternativas tais como LSTM ou ELMo.

Em relação ao desenvolvimento de modelos para a tarefa de classificação de textos, BERT pode ser explorado tanto como um *embedding*, quanto como um modelo principal. Em ambos, os modelos BERT pré-treinados devem ser ajustados ao problema, através de uma etapa de treinamento para ajuste fino usando o conjunto de dados de domínio. Na prática, a diferença entre esses formatos consiste na parametrização definida para a camada de saída da rede, que pode determinar o retorno do *embedding* sobre os dados de entrada, ou as probabilidades para cada classe ao incluir uma camada final com uma função ativação apropriada para o problema de classificação (e.g *sigmoid* ou *softmax*).

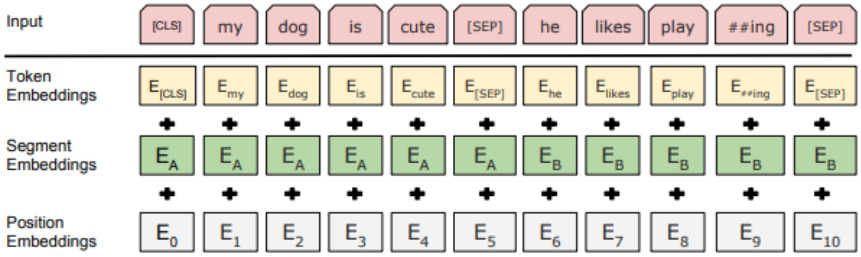
Para usar um modelo pré-treinado BERT, é necessário preparar os dados de entrada segundo o formato específico determinado, como ilustra a Figura 2.4. Nesse formato, o texto deve ser segmentados em *tokens*, cuja sequencia não deve ultrapassar 512 termos. As sequencias devem ter tamanhos iguais e, para cada instância de dados, o início da sentença deve ser sinalizado por uma marcação usando a *tag [CLS]*. As sentenças seguintes devem ser separadas usando a tag *[SEP]*. Além da sequencia de entrada (*input*), a Figura 2.4 apresenta os seguintes componentes:

- *Token Embeddings*: responsável por transformar cada palavra em representações vetoriais de tamanho fixo (768 dimensões);
- *Segment Embeddings*: responsável por identificar as sequencias de entrada conforme separação atribuída usando as tags *[CLS]* e *[SEP]*;
- *Position Embeddings*: camada capaz de atribuir vetores de representação diferentes para um mesmo *token*, baseado em informações fornecidas pelo mecanismo de atenção. Esse mecanismo retém informações relevantes durante as etapas de treinamento, as quais permitem por exemplo, identificar quais contextos o termo em análise já foi empregado, auxiliando a decidir a melhor representação para o termo no seguimento analisado.

Vários modelos pré-treinados BERT apresentam variações principalmente quanto à topologia e dados usados para o desenvolvimento desses modelos. Neste trabalho foi adotado o BERT Base *Uncased*[6] para comparação ao *DAC Stacking* com soluções estado

[6]Disponível em: https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1

Figura 2.4: Representação do formato para entrada de dados BERT



Fonte: Devlin et al. (2019)

da arte em PLN.

2.5 *Ensembles*

Os comitês de classificadores (*ensembles*) são uma técnica para ganho de performance em diferentes tarefas de aprendizado de máquina. A técnica consiste em combinar modelos diferentes visando uma decisão coletiva para a tarefa de classificação ou regressão. A premissa é que cada modelo contribui com um espaço de hipóteses, linguagem de representação e função de avaliação de hipóteses diferentes. Quando agregados, esses modelos trazem à solução maior robustez, pois o espaço de hipóteses gerado pelo modelo final agora considera ótimos mais próximos dos globais, e reduz o custo computacional em treinar um único modelo para uma tarefa complexa (GAMA et al., 2011).

Na formação de um *ensemble* é necessário garantir que os modelos usados, denominados classificadores de base ou classificadores fracos, atendam os seguintes requisitos (ZHANG; MA, 2012): (1) diversidade, pois devem ser capazes de realizar análises independentemente dos demais, gerando resultados que podem ser diferentes para uma mesma instância de dados; e (2) acurácia, onde a taxa de erros dos modelos de base deve ser inferior a 50%.

De acordo com o método empregado para combinar os modelos de base, os *ensembles* podem ser classificados como (ZHANG; MA, 2012):

- *Homogêneos*: usam modelos gerados por um único algoritmo. Nesse caso, diferentes métodos baseados em amostragem de dados de treinamento são empregados para gerar diversidade entre os modelos de base. As abordagens mais usadas para gerar variabilidade a partir dos dados de treinamento são o *Bagging (Bootstrap Aggregation)* e o *Boosting*;

- *Heterogêneos*: emprega diferentes algoritmos de classificação para gerar os mode-

los de base, como meio de incluir diversidade para esse conjunto de classificadores. Esses algoritmos são treinados e testados usando o mesmo conjunto de dados. A combinação desses modelos de base pode ser realizada, segundo duas abordagens principais: pilha e cascata.

Na Generalização em Pilha, ou *Stacking*, a topologia organiza subconjuntos de modelos em níveis. A camada inicial da topologia, denominada nível 0 ou inferior, é formada por classificadores de base que recebem os dados de entrada para treinamento. Cada camada adicional na pilha de classificadores, recebe como entrada as predições da camada anterior, e propagam suas predições para camada seguinte. A última camada dessa pilha, denominada *meta-learner*, é responsável por emitir a predição final para a instância de dados processada, gerenciando as predições dos classificadores de base com uso de algum método (e.g votação majoritária).

Uma variação do *stacking* é a Generalização em Cascata, ou *Cascading*, na qual a cada nível da pilha, novos atributos são gerados com base nas predições dos classificadores correspondentes.

Por fim, a definição dos modelos *ensemble* incluem a seleção de uma abordagem de meta-aprendizado para gerenciar as probabilidades dos classificadores de base e emitir a predição final para uma instância de dados. Entre as abordagens possíveis, destacam-se a seriação (e.g valor máximo), a votação ou os meta-classificadores. No último caso, o modelo é gerado a partir dos metadados, com base nas predições dos classificadores de base.

Neste trabalho, a técnica *ensemble* foi usada para a classificação das condições mentais alvo deste estudo. Para formação do modelo *ensemble*, adotou-se a abordagem de generalização em pilha, segundo dois níveis. No nível inferior, diferentes combinações (topologias) foram testadas para os classificadores de base responsáveis por analisar as amostras e emitir uma probabilidade para cada classe. No nível superior, a abordagem de meta-classificadores foi empregada para ponderar as predições dos classificadores de base, gerando a predição final para uma instância de dados.

2.6 Métricas de Avaliação de Performance

O emprego de algoritmos de aprendizado de máquina para solucionar problemas reais envolve um processo de experimentação, no qual esse modelos são ajustados aos

dados que representam o domínio do problema por sucessivas etapas envolvendo parametrização e treinamento desses modelos. Assim, é de suma importância selecionar métodos adequados para a avaliação de performance dos modelos gerados, permitindo comparar o impacto do uso de algoritmos diferentes para a mesma tarefa, ou mesmo, avaliação de parametrizações diferentes considerando o mesmo algoritmo.

Um modelo de aprendizado de máquina pode ser avaliado segundo diferentes aspectos, tais como acurácia, taxa de erro, compreendibilidade do conhecimento extraído, tempo de aprendizado e requisitos de armazenamento do modelo (GAMA et al., 2011). Considerando as tarefas de classificação ou preditivas, a seleção das métricas está relacionada ao formato de rotulação usado pelos modelos preditivos. A seguir são destacadas as principais métricas empregadas para avaliação de performance em tarefas de classificação envolvendo dois ou mais rótulos de classe, bem como um método para compreensão do conhecimento extraído por modelos de aprendizado profundo.

2.6.1 Métricas para problemas de classificação de rótulo único

Em problemas de classificação de rótulo único, somente um rótulo pode ser atribuído a cada instância do conjunto de dados. Desse modo, a classificação está correta quanto o valor do rótulo previsto para a instância de dados corresponde ao rótulo esperado, e incorreto em caso contrário. Uma matriz de confusão representa os erros e acertos, como ilustra a Tabela 2.3 para um problema binário. A partir da matriz de confusão, importantes métricas são definidas. Desse conjunto de métricas, as mais comumente usadas para avaliação de performance de modelos preditivos de rótulo único são (GAMA et al., 2011):

- *Acurácia*: corresponde à taxa total de acertos da previsão, considerando o conjunto total de instâncias. É uma métrica que avalia o modelo como um todo;

- *Precisão*: para uma dada classe específica, essa medida representa a quantidade de instâncias que foram corretamente classificadas como pertencentes a essa classe, como apresentado pela Equação 2.1;

$$P(C_i) = \frac{VP_{Ci}}{VP_{Ci} + FP_{Ci}} \tag{2.1}$$

- *Revocação*: dentre todas as instâncias de uma classe específica, essa métrica repre-

senta aquelas classificadas corretamente, conforme descrito na Equação 2.2;

$$R(C_i) = \frac{VP_{Ci}}{VP_{Ci} + FN_{Ci}} \tag{2.2}$$

- *Taxa de Falso Positivo (TFP)*: corresponde à taxa de erro considerando uma determinada classe (Equação 2.3);

$$TFP(C_i) = \frac{FP_{Ci}}{FP_{Ci} + VN_{Ci}} \tag{2.3}$$

- *Medida F*: combina as métricas de precisão e revocação usando média harmônica ponderada. Se a análise é realizada considerando a precisão e revocação igualmente importantes, essa medida é denominada F1 (Equação 2.4).

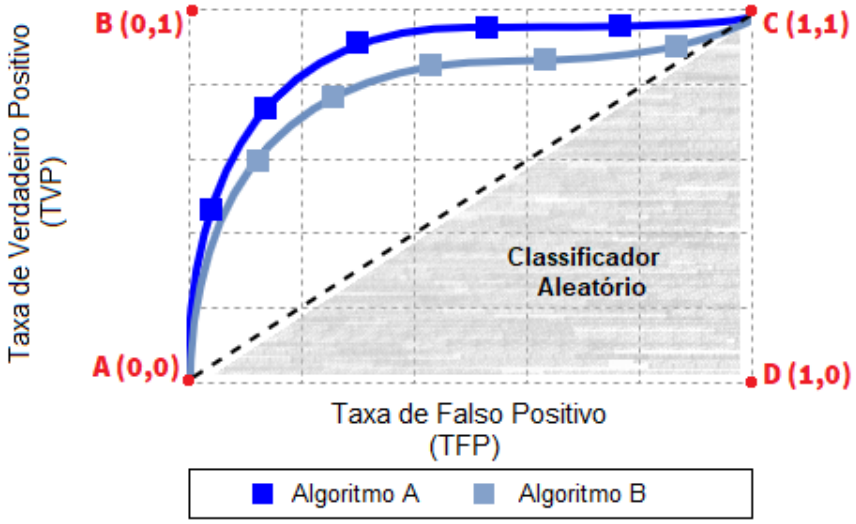
$$F1(C_i) = \frac{2 \times P(C_i) \times R(C_i)}{P(C_i) + R(C_i)} \tag{2.4}$$

Tabela 2.3: *Matriz de Confusão*: Exemplo para problema de classificação binária

		Valor Previsto	
		Classe 1	**Classe 2**
Valor Real	**Classe 1**	$\boldsymbol{C_1}$ Verdadeiro Positivo (VP)	$\boldsymbol{C_2}$ Falso Positivo (FP)
	Classe 2	$\boldsymbol{C_1}$ Falso Negativo (FN)	$\boldsymbol{C_2}$ Verdadeiro Negativo (VN)

Outra métrica frequentemente usada para avaliação de classificadores binários é a Curva Receiving Operating Characteristics (ROC) (GAMA et al., 2011). O gráfico da curva ROC é gerado no plano cartesiano, onde as taxa de falso positivo (TFP) e verdadeiro positivo (TVP) correspondem aos eixos X e Y, respectivamente. A Figura 2.5 destaca os principais pontos desse gráfico. Um classificador perfeito é representado pelo ponto B de coordenada (0,1). De modo oposto, um classificador com erro máximo é representado pelo ponto D (1,0). Os pontos A (0,0) e C (1,1) representam modelos que predizem todas as amostras como negativas e positivas, respectivamente. Por fim, curvas abaixo da reta $\overline{AC}$, representam classificadores totalmente aleatórios. Além do gráfico, a medida de área abaixo da curva ROC, Area Under Curve (AUC), também é usada para avaliação dos modelos. Essa medida apresenta valores entre 0 e 1, sendo que os mais próximos de 1 indicam melhor performance.

Figura 2.5: Curva ROC: Exemplo de comparativo entre dois algoritmos



Fonte: Adaptado de Gama et al. (2011)

2.6.2 Métricas para problemas de classificação multirrótulo

Na classificação multirrótulo, cada instância do conjunto de dados pode ser associada a mais de um rótulo. A análise de performance dos modelos de classificação multirrótulo pode ser considerada total ou parcialmente correta, ou ainda, totalmente incorreta, segundo a quantidade de rótulos atribuídos para cada instância dos dados. As métricas tradicionais podem ser usadas para análise de problemas multirrótulo, considerando cada classe individualmente. No entanto, esses métodos não consideram as correlações entre as diferentes classes.

Para capturar a noção de parcialmente correto, uma estratégia é avaliar a diferença média entre os rótulos previstos e os rótulos reais para cada exemplo de teste e, em seguida, a média de todos os exemplos no conjunto de teste. Essa abordagem é chamada de avaliações baseadas em exemplos. Nessa categoria de abordagens destacam-se as métricas (GAMA et al., 2011):

a) *Exact Match Ratio (EMR)*: é uma métrica rígida de avaliação, na qual uma amostra é pontuada como corretamente classificada se todas as classes possíveis estiverem corretamente identificadas. Essa métrica é calculada conforme Equação (2.5);

$$EMR = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Z_i) \tag{2.5}$$

b) *Hamming Loss (HL)*: é uma métrica de avaliação baseada em função de *ranking*,

a qual usa como medida de ordenação a distância de Hamming entre os valores reais e preditos, ponderando tanto o erro de predição (um rótulo incorreto é atribuído), quanto o erro de perda (um rótulo deixou de ser predito) (Equação 2.6). Considerada uma medida mais branda, essa métrica reporta quantas vezes, em média, foram registradas perdas em cada classes. Valores menores para essa métrica indicam melhor performance do modelo;

$$HL(\widehat{f},X) = \frac{1}{n}\sum_{i=1}^{n}\frac{a(Y_i,Z_i)}{k} \tag{2.6}$$

c) *Acurácia, Precisão, Revocação e Medida F*: na abordagem baseada em exemplos, essas métricas podem ser adaptadas para fornecer uma medida parcial de performance, considerando cada classe em relação a média de todas as instância. Algumas adaptações incluem um critério de penalização de erros cometidos na predição das classes.

2.6.3 Compreensão de modelos profundos usando SHAP

Os algoritmos de aprendizado profundo são considerados modelos não interpretáveis devido as suas previsões não serem claramente justificáveis com regras ou fluxos de tomada de decisão. Para compreender os critérios destes modelos ao realizar as previsões, diferentes métodos são propostos. De modo geral, esses métodos geram variações nos dados de entrada, com base em algoritmo para seleção de candidatos ótimos, e avaliam o impacto dessas alterações na previsão realizada pelo modelo. Entre esses métodos, destaca-se o SHAP (LUNDBERG; LEE, 2017).

No método SHAP, o algoritmo mitiga variações para os dados de entrada, escolhendo os melhores candidatos segundo a teoria dos jogos, e analisa o impacto dessas variações na classificação de uma amostra. A previsão é razoavelmente distribuída entre os valores Shapley (HART, 1989) calculados para as *features*, fornecendo explicações contrastivas que comparam as previsões em relação à previsão média para a instância. Ao final do processo, para cada instância analisada, o método fornece uma lista com as *features* que mais influenciaram, tanto positiva, quanto negativamente para a decisão de classificação do modelo. Entre as diferentes abordagens de estimativa de resultados SHAP, destaca-se a baseada em explanações do *kernel* usada por este trabalho para análise das arquiteturas de aprendizado profundo via implementação disponível na biblioteca

Kernel Explainer[7].

Os valores Shapley podem ser usados tanto para a realização de uma explicação local, quanto global. Neste caso, os valores dos coeficientes das diferentes amostras são agregados. Este trabalho usa os valores SHAP gerados para os termos que compõem uma amostra como um índice para análise de sua importância em relação à classificação de uma instância.

2.7 *Framework Ray Tune*

Tabela 2.4: *Ray Tune*: Principais parâmetros e suas configurações.

Função	Parâmetro	Descrição	Valor Definido
Tune	scheduler	Função usada para determina os valores ótimos para seleção de parâmetros de treinamento.	Async HyperBand Scheduler*
	stop	Critério de parada para o treinamento do modelo (somente a opção acurácia média).	mean accuracy
		Definie o valor de "paciência", isto é, quantas iteração devem ser executadas após atiginr o critério de parada.	50 iteration
	num_samples	Total de testes realizados (modelos gerados).	10
	resource_per_trial	Número de recursos alocados por tentativa (total de CPUs e/ou GPUs)	4
Async HyperBand Scheduler	time_attr	Medida de progresso para o treinamento dos modelos.	training iteraction
	metric	Métrica de performance usada como critério de parada para o treinamento dos modelos. Deve ser a mesma definida na função Tune.	mean accuracy
	mode	Determina se o critério de parada deve considerar o valor máximo ou mínimo para a métrica de monitorada.	max
	max_t	Determina o número máximo de iterações para o treinamento do modelo.	200

*Esse valor definide o uso do algoritmo ASHA para seleção de valores ótimos (LI et al., 2020).

O *framework Ray Tune*[8] engloba um robusto conjunto de bibliotecas que auxiliam no processo de treinamento de modelos de aprendizado de máquina. Essas ferramentas podem ser usadas tanto para paralelizar o processo de treinamento, quanto explorar valores para um conjunto de parâmetros pré-definidos para ajuste do modelo explorado.

Entre as ferramentas disponibilizadas por esse *framework*, destaca-se a funcionalidade *Tune*, a qual apresenta diferentes algoritmos para a heurística de seleção de valores ótimos de cada parâmetro informado na etapa de treinamento. Essa função realiza o treinamento dos modelos de modo assíncrono, variando os parâmetros durante iterações ao longo desse processo. A eficácia dessa abordagem consiste em interromper treinamen-

[7]https://shap.readthedocs.io/en/latest/#shap.KernelExplainer

[8]Disponível em: https://docs.ray.io/en/latest/ray-overview/index.html

tos cuja parametrização não apresenta ganho de performance ao longo das iterações, bem como o ajuste de parâmetros durante o processo de treinamento.

Neste trabalho, a funcionalidade *Tune* foi usada para treinamento dos classificadores de base dedicados à função de diferenciadores, como uma abordagem adicional para ajuste fino desses modelos. A Tabela 2.4 destaca as principais configurações definidas para o treinamento dos modelos usando a função *Tune*.

3 TRABALHOS RELACIONADOS

Este capítulo se dedica a elucidar o contexto em que essa dissertação está inserida, apresentando uma visão geral e trabalhos relacionados que exploram o potencial das técnicas de aprendizado de máquina para a identificação de transtornos mentais a partir do conteúdo de mídias sociais. Em seguida, são apresentados trabalhos que focam no estudo dos transtornos de ansiedade, depressão e na condição de comorbidade entre essas desordens. Por fim, destacamos as técnicas de aprendizado profundo que apresentaram bons desempenhos para classificação de usuários no contexto dos transtornos alvo deste estudo.

3.1 Visão Geral: Aprendizado de Máquina aplicado à Saúde Mental e Bem-estar

Uma revisão sistemática (WONGKOBLAP; VADILLO; CURCIN, 2017) mostra que muitos estudos têm concentrado esforços na extração de padrões a partir de textos e interações disponíveis nas mídias sociais, com a finalidade de compreender comportamentos e gerar modelos preditivos no domínio da saúde mental. Esses trabalhos se distinguem nos seguintes aspectos, detalhados nesta seção: (a) *rede social estudada*; (b) *técnicas de aprendizado de máquina empregadas*; (c) *propósito em relação aos cuidados com bem estar e saúde mental*; (d) *objetivo de análise (usuários ou postagens)*; e (e) *transtornos mentais estudados*.

Em relação às redes sociais, destacam-se o Twitter (CHOUDHURY; COUNTS; HORVITZ, 2013; PARK; MCDONALD; CHA, 2013; COPPERSMITH; DREDZE; HARMAN, 2014; COPPERSMITH; HARMAN; DREDZE, 2014; TSUGAWA et al., 2015; PREOŢIUC-PIETRO et al., 2015; COPPERSMITH et al., 2015; Keumhee Kang; Chanhee Yoon; Eun Yi Kim, 2016; CHOUDHURY et al., 2016a; CHOUDHURY et al., 2017; LOVEYS et al., 2017; BENTON; MITCHELL; HOVY, 2017; DUTTA; MA; CHOUDHURY, 2018; GRUDA; HASAN, 2019) e o Facebook[1] (CHOUDHURY et al., 2014; PARK et al., 2015; ISLAM et al., 2018) por sua abrangência mundial e por apresentarem tema livre para discussão. O Reddit também tem sido explorado por permitir que usuários participem de modo anônimo e criem canais para discussões sobre temas específicos, sem apresentar limite de palavras para publicação (CHOUDHURY et al., 2016b; BAGROY; KUMARAGURU; CHOUDHURY, 2017; GKOTSIS et al., 2017; SHEN;

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

RUDZICZ, 2017; YATES; COHAN; GOHARIAN, 2017; IRELAND; ISERMAN, 2018; SHARMA; CHOUDHURY, 2018; IVE et al., 2018; COHAN et al., 2018; CACHEDA et al., 2019; MATERO et al., 2019; TADESSE et al., 2019a). Outras fontes de dados utilizadas incluem o Instagram (MANIKONDA; CHOUDHURY, 2017; MANN; PAES; MATSUSHIMA, 2020) e serviços de ajuda às pessoas que sofrem com algum transtorno mental oferecidos por governos locais (ALTHOFF; CLARK; LESKOVEC, 2016; YATES; COHAN; GOHARIAN, 2017).

Em termos de técnicas de aprendizado de máquina, grande parte dos trabalhos desenvolvidos até o momento aplicam abordagens de aprendizado raso juntamente com uma extensa engenharia de *features* para identificação de transtornos mentais. Uma compilação desses trabalhos é discutida na Seção 3.3. Recentemente, os benefícios dos modelos de aprendizado profundo têm sido explorados para solução dessa categoria de problemas. Sua principal vantagem em relação às técnicas de aprendizado raso é o aprendizado de padrões mais complexos que minimizam, ou mesmo dispensam a etapa de engenharia de *features*. Os trabalhos relacionados que exploram aprendizado profundo são detalhados na Seção 3.4.

Quanto ao propósito, os estudos realizados apresentam diferentes aspectos relacionados aos cuidados com a saúde mental e bem estar. Entre os temas abordados estão a identificação de traços de aconselhamentos mais bem sucedidos em fóruns de apoio à saúde mental (ALTHOFF; CLARK; LESKOVEC, 2016; SHARMA; CHOUDHURY, 2018), ou a análise de distanciamento psicológico mediante situações de violência e discriminação (CHOUDHURY et al., 2016a). Contudo, a maioria dos trabalhos foca no desenvolvimento de um modelo preditivo para transtornos mentais, onde a classificação de um transtorno específico é prevalente, como em (CHOUDHURY et al., 2014; LIN et al., 2014; COPPERSMITH; HARMAN; DREDZE, 2014; Keumhee Kang; Chanhee Yoon; Eun Yi Kim, 2016; SHEN; RUDZICZ, 2017).

Em relação ao objetivo de análise, a abordagem preditiva pode ser proposta com intuito de classificar as mensagens, ou os usuários. Na classificação de mensagens ou *posts*, o modelo prediz a classe de cada mensagem, sem considerar a qual usuário pertence. Essa abordagem apresenta diferentes aplicações, tais como a moderação de conteúdos para serviços *online* de suporte à saúde mental (SHARMA; CHOUDHURY, 2018). Já na segunda abordagem, o foco é predizer a saúde mental de cada usuário com base em seu conjunto de mensagens. Essa tarefa apresenta maior complexidade computacional devido à natureza dos transtornos mentais, onde os traços da doença não são manifestos

de modo contínuo e regular (American Psychiatric Association, 2013). Assim, o método deve ser capaz de predizer um usuário diagnosticado, ainda que nem todas as suas mensagens contenham traços do transtorno mental.

Finalmente, em relação aos transtornos mentais, a depressão é o mais estudado (PARK; MCDONALD; CHA, 2013; TSUGAWA et al., 2015; PARK et al., 2015; Keumhee Kang; Chanhee Yoon; Eun Yi Kim, 2016; Hu et al., 2015; ISLAM et al., 2018; CACHEDA et al., 2019; TADESSE et al., 2019a; MANN; PAES; MATSUSHIMA, 2020), seguido de ideação suicida (CHOUDHURY et al., 2016b; CHOUDHURY et al., 2017; MATERO et al., 2019) e do Transtorno de Estresse Pós-Traumático (TEPT) (CHOUDHURY; COUNTS; HORVITZ, 2013; COPPERSMITH; HARMAN; DREDZE, 2014; PREOŢIUC-PIETRO et al., 2015). Poucos trabalhos abordam a ansiedade e sua comorbidade com a depressão (BENTON; MITCHELL; HOVY, 2017; COHAN et al., 2018).

Neste trabalho, os esforços de pesquisa concentram-se no estudo dos transtornos de ansiedade, depressão e sua condição de comorbidade para a classificação de usuários, segundo dados extraídos da rede social Reddit. Assim, as seções seguintes são dedicadas à apresentação dos principais trabalhos realizados com foco na identificação dessas desordens a partir de conteúdos de mídias sociais.

3.2 Conjuntos de Dados para Geração de Modelos de Aprendizado de Máquina

A capacidade de generalização de modelos preditivos baseados em técnicas de aprendizado de máquina supervisionado depende fortemente da disponibilidade de conjuntos de dados de treinamento grandes e sem viés. Em especial, considerando a classificação automática de transtornos mentais, a disponibilidade de conjuntos de dados de treinamento com essas características é um fator crítico. Isso porque, os estudos desenvolvidos pela área de saúde mental, embora apresentem alta confiabilidade para os rótulos das instâncias, tipicamente resultam em pequenos conjuntos de dados devido ao baixo número de participantes.

A formação de conjuntos de dados a partir de conteúdos extraídos das redes sociais surge como uma alternativa para apoiar o desenvolvimento de modelos computacionais dedicados à identificação automática de transtornos mentais. De modo geral, o processo de formação desses conjuntos pode ser sintetizado nas seguintes etapas: (1) definição do conjunto de expressões chave para busca de usuários diagnosticados e/ou auto-declarados;

(2) validação desse subconjunto de dados quanto à veracidade das declarações de auto-diagnóstico dos usuários; (3) definição de critérios para formação dos grupos de usuários de controle ou saudáveis; e (4) coleta do histórico de postagens dos usuários que atendem aos critérios definidos para os filtros de pesquisa. Conforme a estrutura e os recursos disponíveis em cada rede social, esse processo apresenta algumas variações para a etapa de busca de usuários, as quais serão discutidas a seguir.

Na plataforma Reddit, a comunicação dos usuários está organizada segundo comunidades, denominadas *subreddits*, que são dedicadas à discussão de um tema específico. Sua estrutura propicia tema livre para discussões sem restringir o tamanho das postagens dos usuários. Essas características são vistas como um meio de incentivar as expressões dos usuários, sendo destacada como uma vantagem para muitos estudos em saúde mental (BAGROY; KUMARAGURU; CHOUDHURY, 2017; YATES; COHAN; GOHARIAN, 2017; COHAN et al., 2018; CACHEDA et al., 2019; TADESSE et al., 2019b). Esses estudos definem a busca de usuários diagnosticados através da identificação de sentenças de auto-declaração em *subreddits* dedicados aos temas relacionados aos transtornos mentais. As sentenças de auto-declaração incluem afirmações como "Eu fui diagnosticado com". O grupo de controle é composto tipicamente por usuários que não apresentam postagens em *subreddits* dedicados à saúde mental. A maioria dos trabalhos realiza a coleta de dados usando a Application Program Interface (API) do Reddit (CHOUDHURY et al., 2016b; BAGROY; KUMARAGURU; CHOUDHURY, 2017; COHAN et al., 2018; CACHEDA et al., 2019; TADESSE et al., 2019b). Uma abordagem diferente é proposta em Bagroy, Kumaraguru e Choudhury (2017), onde a coleta de dados inicia pelo uso da ferramenta de consulta Google's BigQuery e a API do Reddit é usada para complementar os dados dos usuários selecionados.

Nos estudos envolvendo a rede social Twitter, os trabalhos definem um grupo de palavras-chave ou sentenças de auto-declaração que melhor representam o transtorno alvo da pesquisa e, com base nesse conjunto de palavras, recuperam *tweets* usando a API do Twitter (COPPERSMITH; DREDZE; HARMAN, 2014; COPPERSMITH et al., 2015; GRUDA; HASAN, 2019). Para a formação do conjunto de controle, são selecionados aleatoriamente usuários cujos *tweets* não contenham as palavras ou sentenças usadas para definir o grupo de diagnóstico. Por fim, para os usuários que atendem aos requisitos definidos para a pesquisa, o histórico de postagens é recuperado.

O recrutamento de voluntários é uma prática comum em trabalhos que utilizam dados do Facebook (CHOUDHURY et al., 2014; PARK et al., 2015), do Instagram (MA-

NIKONDA; CHOUDHURY, 2017; MANN; PAES; MATSUSHIMA, 2020) e, em menor proporção, do Twitter (TSUGAWA et al., 2015; PARK; MCDONALD; CHA, 2013). Nesses casos, os participantes são submetidos a uma avaliação psicométrica via preenchimento de questionários, tais como o Center for Epidemiologic Studies Depression Scale (CES-D), que permite quantificar a presença de um transtorno mental em níveis de intensidade. Com base nos resultados obtidos com essa avaliação, os autores definem os participantes em usuários saudáveis ou diagnosticados. Ao final dessa etapa, o histórico de postagens de cada usuário é recuperado usando a API oficial de cada plataforma. No Instagram, tanto o conteúdo textual composto de legenda, comentários e *hashtags*, quanto as fotos podem ser capturadas.

O processo de rotulagem das instâncias é um dos grandes desafios para a formação dos conjuntos de dados usados para a classificação de transtornos mentais. Em trabalhos que usam a abordagem de recrutamento de voluntários para formação do conjunto de dados (PARK; MCDONALD; CHA, 2013), essa dificuldade é sanada, uma vez que os usuários são classificados conforme escore obtido por um instrumento de avaliação clínica. No entanto, por depender de voluntários, esses estudos costumam apresentar uma quantidade de dados muito menor em termos do total de indivíduos se comparados a outros trabalhos que utilizam métodos automáticos para a rotulagem de dados, tais como (COHAN et al., 2018).

Nas abordagens de rotulagem automática, é necessário eliminar casos de falso positivo, onde afirmações de auto-diagnóstico podem ser declaradas com o sentido de brincadeira ou ironia por usuários saudáveis. Os trabalhos relacionados executam esta etapa usando as seguintes estratégias: (a) inspeção manual realizada pelos próprios pesquisadores com auxílio de especialistas, tais como (CHOUDHURY et al., 2016b; YATES; COHAN; GOHARIAN, 2017); (b) via uso de abordagens para rotulagem ou correspondência de diagnóstico, acompanhadas por inspeção manual para uma subamostragem de dados (CHOUDHURY et al., 2017; COHAN et al., 2018); ou (c) através de ferramentas de recrutamento de voluntários para validação manual (CHOUDHURY; COUNTS; HORVITZ, 2013).

Por fim, outro desafio é garantir que os dados não contenham manifestações que resultem em viés para a tarefa de classificação. Para tanto, é necessário eliminar expressões óbvias que tornem artificialmente fácil a identificação de usuários diagnosticados. Muitos trabalhos informam que excluem as postagens originais usadas para identificar os usuários diagnosticados (PARK; MCDONALD; CHA, 2013; COPPERSMITH;

DREDZE; HARMAN, 2014; Keumhee Kang; Chanhee Yoon; Eun Yi Kim, 2016). Outros mencionam que o método não contemplou uma busca ampla para remoção de demais expressões que permitam facilmente associar os usuários à classe de diagnosticados (COPPERSMITH et al., 2015). A abordagem mais rigorosa inclui uma etapa de remoção de postagens, a partir da identificação de um conjunto de expressões e termos definidos como mais usados, segundo o perfil de usuário diagnosticado (IRELAND; ISERMAN, 2018; YATES; COHAN; GOHARIAN, 2017; COHAN et al., 2018).

Este trabalho utiliza o conjunto de dados Self-reported Mental Health Diagnoses (SMHD) (COHAN et al., 2018), formado pela extração do conteúdo de usuários da rede social Reddit, auto-declarados como diagnosticados segundo nove transtornos mentais. O processo de formação desse conjunto de dados, bem como seu uso por este trabalho são descritos em detalhes no Capítulo 4.

3.3 Aprendizado Raso na Detecção de Transtornos Mentais

Muitos trabalhos contribuíram para a caracterização de transtornos mentais a partir de conteúdo textual das mídias sociais. O reconhecimento dos padrões relacionados às desordens foi amplamente explorado usando abordagens supervisionadas que combinam aprendizado raso com extensa engenharia de *features*.

Uma recente revisão sistemática sobre pesquisas na área computacional com foco em depressão (GIUNTINI et al., 2020), destaca que esse transtorno foi muito estudado, tanto isoladamente (TSUGAWA et al., 2015; Keumhee Kang; Chanhee Yoon; Eun Yi Kim, 2016; Hu et al., 2015; ISLAM et al., 2018; CACHEDA et al., 2019; TADESSE et al., 2019b), quanto sob diferentes contextos, tais como solidão (PARK et al., 2015), ideação suicida (CHOUDHURY et al., 2017; CHOUDHURY et al., 2016b) e pós-parto (CHOUDHURY; COUNTS; HORVITZ, 2013; CHOUDHURY et al., 2014)). Para caracterização dessa desordem, os algoritmos mais usados foram SVM (Keumhee Kang; Chanhee Yoon; Eun Yi Kim, 2016; TSUGAWA et al., 2015; CHOUDHURY; COUNTS; HORVITZ, 2013), seguido de regressão logística (Hu et al., 2015; CHOUDHURY et al., 2014; CHOUDHURY et al., 2016b). Alguns trabalhos usaram abordagens baseadas em árvores (ISLAM et al., 2018; CACHEDA et al., 2019) e estatísticas (PARK et al., 2015; CHOUDHURY et al., 2017).

Poucos trabalhos abordam os transtornos de ansiedade. Alguns estudos propõem modelos preditivos baseado no algoritmo SVM (DUTTA; MA; CHOUDHURY, 2018) e

Bayesian Ridge Regression (GRUDA; HASAN, 2019). Outros trabalhos destacam *insights* sobre a desordem como sua principal contribuição. Ireland e Iserman (2018) obtêm *insights* estabelecendo comparativos entre as postagens neutras e as identificadas como contendo traços de ansiedade, bem como entre conteúdos gerados por usuários ansiosos, comparando suas mensagens quando postadas em fóruns de ansiedade e em fóruns de outros temas na rede social Reddit. Os autores exploram algoritmos de regressão logística e árvores de decisão combinados com categorias de palavras extraídas do dicionário LIWC. O potencial de micro padrões de comunicação na rede Twitter para identificação de traços de ansiedade é explorada em (LOVEYS et al., 2017) através de análise de sentimentos usando a ferramenta VADER (HUTTO; GILBERT, 2014).

Alguns trabalhos abrangem a classificação de múltiplas desordens, incluindo depressão (COPPERSMITH; DREDZE; HARMAN, 2014; PREOŢIUC-PIETRO et al., 2015), e depressão e ansiedade (COPPERSMITH et al., 2015; BAGROY; KUMARAGURU; CHOUDHURY, 2017; PRIMACK et al., 2017). Embora apresentem *insights* sobre padrões semelhantes encontrados em mais de um transtorno, esses estudos não focam no estudo de traços ou padrões que podem identificar a comorbidade ou mesmo distinguir essa condição das manifestações individuais dos transtornos envolvidos.

Em comum, todos os trabalhos supracitados combinam uma abordagem de aprendizado raso com uma extensa engenharia de *features* para a exploração de informações distintas em redes sociais. As informações mais exploradas estão resumidas abaixo:

- *textos de postagens*: usados em todos os trabalhos, diferenciam-se quanto à forma de representação. Entre as mais comuns estão as variações de *n-grams* (COPPERSMITH; DREDZE; HARMAN, 2014; COPPERSMITH; HARMAN; DREDZE, 2014; CHOUDHURY et al., 2016b; CHOUDHURY et al., 2017; BAGROY; KUMARAGURU; CHOUDHURY, 2017; TADESSE et al., 2019b), *Bag of Words* (TSUGAWA et al., 2015; CACHEDA et al., 2019) e frequência de palavras (CHOUDHURY; COUNTS; HORVITZ, 2013; CHOUDHURY et al., 2014; COPPERSMITH et al., 2015; PREOŢIUC-PIETRO et al., 2015; CHOUDHURY et al., 2016b; CHOUDHURY et al., 2017; DUTTA; MA; CHOUDHURY, 2018; ISLAM et al., 2018; CACHEDA et al., 2019);

- *estruturas morfológicas*: muitos trabalhos extraem informações adicionais de texto, sendo comuns o uso do tempo verbal e a pessoa (verbos), gênero (substantivos) e a classe gramatical das palavras. A extração dessas representações é realizada a partir de dicionários como LIWC (CHOUDHURY; COUNTS; HORVITZ, 2013;

CHOUDHURY et al., 2014; COPPERSMITH; DREDZE; HARMAN, 2014; COPPERSMITH et al., 2015; PREOŢIUC-PIETRO et al., 2015; CHOUDHURY et al., 2016b; CHOUDHURY et al., 2017; DUTTA; MA; CHOUDHURY, 2018; ISLAM et al., 2018; TADESSE et al., 2019b) ou WordNet (Keumhee Kang; Chanhee Yoon; Eun Yi Kim, 2016);

- *sociais*: são definidas conforme cada plataforma de mídia social (Twitter, Facebook, Reddit). No entanto, algumas *features* sociais são comuns a todas as plataformas e amplamente usadas, tais como diferentes estatísticas relacionadas ao número de postagens (CHOUDHURY; COUNTS; HORVITZ, 2013; CHOUDHURY et al., 2014; TSUGAWA et al., 2015), *replies* (CHOUDHURY; COUNTS; HORVITZ, 2013; DUTTA; MA; CHOUDHURY, 2018) e *likes* (CHOUDHURY et al., 2014; PARK et al., 2015). Alguns trabalhos exploram outras *features* específicas à plataforma, tais como número de *retweets*, *followings* e *followers* (CHOUDHURY; COUNTS; HORVITZ, 2013; TSUGAWA et al., 2015);

- *sentimento*: englobam o uso de abordagens que permitem extrair, a partir do texto, o sentimento próprio ao domínio, relacionando termos a categorias de interesse do domínio. A abordagem mais usada é o LIWC (CHOUDHURY; COUNTS; HORVITZ, 2013; COPPERSMITH; DREDZE; HARMAN, 2014; COPPERSMITH et al., 2015; IRELAND; ISERMAN, 2018; ISLAM et al., 2018). A ferramenta VADER é usada em (LOVEYS et al., 2017; DUTTA; MA; CHOUDHURY, 2018), enquanto que a ferramenta ANEW é usada em (CHOUDHURY; COUNTS; HORVITZ, 2013). Alguns trabalhos usaram léxico de domínio gerados a partir do corpus (COPPERSMITH; DREDZE; HARMAN, 2014) e modelo de traços de personalidade *BigFive* (PREOŢIUC-PIETRO et al., 2015);

- *imagens*: consiste em usar os atributos de imagens, tais como brilho, saturação e contraste, para auxiliar na composição do conjunto de características do usuário. Manikonda e Choudhury (2017) utilizam a biblioteca OpenCV[2] para extração dessas *features* a partir de imagens capturadas do Instagram.

Para a redução de dimensionalidade, os trabalhos propõem o uso de LDA (PREOŢIUC-PIETRO et al., 2015; TSUGAWA et al., 2015; CHOUDHURY et al., 2017; TADESSE et al., 2019b), LIWC (IRELAND; ISERMAN, 2018; ISLAM et al., 2018;

[2]https://opencv.org/

SHARMA; CHOUDHURY, 2018) ou Greedy Stepwise (GS) (Hu et al., 2015) para seleção de *features* ou tópicos, e PCA para a redução de *features* (CHOUDHURY; COUNTS; HORVITZ, 2013).

A Tabela 3.1 resume as principais características dos trabalhos descritos nesta seção. Embora esses estudos forneçam contribuições importantes para auxiliar a elucidar ou ratificar padrões e comportamentos observados na literatura psicológica, algumas limitações são observadas. Poucos trabalhos dedicam-se à identificação da ansiedade (COPPERSMITH et al., 2015; CHOUDHURY et al., 2016b; DUTTA; MA; CHOUDHURY, 2018; GRUDA; HASAN, 2019). Dos que o fazem, nenhum explora a condição de comorbidade com o transtorno de depressão. Em comum, os trabalhos discutidos apresentam uma solução específica, baseada em características extraídas considerando o transtorno alvo e a rede social usada. Esse extenso conjunto de *features* tem como limitação a necessidade de ser redefinido ou ajustado em caso de mudança do transtorno mental alvo ou da rede social. Na próxima seção, são apresentados os trabalhos desenvolvidos segundo a abordagem de aprendizado profundo, a qual tem se mostrado eficiente na abstração de padrões mais complexos, dispensando a etapa de engenharia de *features*.

Tabela 3.1: Identificação de Desordens: Abordagem Supervisionada de Aprendizado Raso.

Trabalho	Abordagem de Classificação	Problema de Sáude Mental (Transtorno)	Objetivo do Estudo	Fonte de Dados	Principais Features	Abordagem Preditiva
Ireland e Iserman (2018)	binário, único rótulo	Ansiedade	Classificação de Posts	Reddit	Textual e Sentimento (LIWC)	Decision Tree's
Dutta, Ma e Choudhury (2018)	binário, rótulo único	Ansiedade	Classificação de Usuários	Twitter	Textuais (LIWC), Interações Sociais Sentimento (VADER)	VAR + SVM
Gruda e Hasan (2019)	binário, rótulo único	Ansiedade	Classificação de Usuários	Twitter	Word Embedding Glove Emojis (uni e bi grama)	Bayesian Ridge Regression
Park, McDonald e Cha (2013)	-	Depressão	Análise de Posts	Twitter	Textuais, Sentimento (LIWC)	Testes Estatísticos
Islam et al. (2018)	binário, rótulo único	Depressão	Classificação de Posts	Facebook	Textual, Sentimento e Temporal (LIWC)	Decision Tree's
Tsugawa et al. (2015)	binário, único rótulo	Depressão	Classificação de Usuários	Twitter	Textuais, Sentimento, Interações sociais Léxico de Domínio (CES-D e BDI) Seleção de Tópicos (LDA)	SVM
Keumhee Kang, Chanhee Yoon e Eun Yi Kim (2016)	binário, único rótulo	Depressão	Classificação de Usuários	Twitter	Textuais (WordNet) Sentimento (VSO, SentiStrength) Emoticons e Imagem	SVM
Hu et al. (2015)	binário, rótulo único	Depressão	Classificação de Usuários	Sina Weibo	Textuais (Wen Xin), Interações Sociais Léxico de Domínio (questionário CES-D) Seleção de Características (Greedy Stepwise)	Regressão Logística
Cacheda et al. (2019)	binário, rótulo único	Depressão	Classificação de Usuários	Reddit	Textual (similaridade, frequência de termos) Características Temporais, Ranking Posts	Random Forest
Tadesse et al. (2019a)	binário, rótulo único	Depressão	Classificação de Usuários	Reddit	Textuais (N e bi-grams), Sentimento (LWIC) Redução de Dimensionalidade (LDA)	Multilayer Percepton
Choudhury, Counts e Horvitz (2013)	binário, rótulo único	Depressão e Estresse Pós-parto	Classificação de Usuários	Twitter	Textuais (LIWC), Interação Sociais Sentimento (LIWC e ANEW) Redução de Dimensionalidade (PCA)	SVM
Primack et al. (2017)	-	Depressão e Ansiedade	Análise de Usuários	Todas Plataformas	Inventários de Avaliação Psicométrica Entrevista Semi-estruturada	Regressão Logística
Choudhury et al. (2017)	-	Depressão e Ideação Suicida	Análise de Usuários	Twitter, Reddit	Textuais (LIWC), Interações Sociais Seleção de Tópicos (LDA) Medida de Distanciamento Psicológico	Campos Gaussianos, Funções Harmônicas
Coppersmith, Dredze e Harman (2014)	binário, rótulo único	Depressão e outras 3 condições mentais	Classificação de Usuários	Twitter	Textuais (LM), Sentimento (LIWC) Interações Sociais, Léxico de Domínio	Regressão Log Linear
Park et al. (2015)	binário, rótulo único	Depressão e Solidão	Classificação de Usuários	Facebook	Léxico de Domínio (CES-D e BDI) Interações Sociais	Testes Estatísticos
Preoţiuc-Pietro et al. (2015)	binário, rótulo único	Depressão e TEPT	Classificação de Usuários	Twitter	Textuais (LIWC), Interações Sociais, Léxico de Domínio Modelo de Personalidade Big Five	Regressão Logística
Choudhury et al. (2014)	binário, rótulo único	Depressão Pós-parto	Classificação de Usuários	Facebook	Textuais, Sentimento (LIWC) Interações Sociais	Regressão Logística
Bagroy, Kumaraguru e Choudhury (2017)	binário, rótulo único	Depressão e Ansiedade na forma de Estresse	Classificação de Usuários	Reddit	Textuais (n-grams) Interações Sociais	Regressão Logística
Loveys et al. (2017)	binário, rótulo único	Múltiplas desordens incluindo Ansiedade e Depressão	Classificação de Usuários	Twitter	Suplementação de dados da rede social com dados do site OurDataHelps.org.	Análise de Categorias (VADER)
Choudhury et al. (2016b)	binário, rótulo único	Múltiplas desordens incluindo Ansiedade e Depressão	Classificação de Usuários	Reddit	Textuais (LIWC), Interações Sociais Medida de Distanciamento Psicológico	Regressão Logística
Manikonda e Choudhury (2017)	-	Múltiplas desordens incluindo Ansiedade e Depressão	Análise de Usuários	Instagram	Textuais e Sentimento (LIWC) Hashtags (TwitterLDA), Imagem (OpenCV)	Análises de Relações (features e transtornos)
Coppersmith et al. (2015)	binário, rótulo único	Múltiplas desordens incluindo Ansiedade e Depressão	Classificação de Usuários	Twitter	Textuais, Sentimento (LIWC) Demográficas (idade e gênero)	CLM

3.4 Aprendizado Profundo na Detecção de Transtornos Mentais

Um trabalho pioneiro no uso de aprendizado profundo para a classificação de usuários depressivos é apresentado por Yates, Cohan e Goharian (2017). A partir de um grande conjunto de dados de usuários Reddit auto-diagnosticados, os autores propõem uma abordagem de classificação binária, usando uma arquitetura CNN para processar o conjunto de *posts* de cada usuário. Com a finalidade de alavancar a performance do modelo, os autores experimentam diferentes formas para a definição do conjunto de *posts* de um usuário, variando a ordem, o número de termos por postagem e o total de postagens por usuário. Os melhores resultados (*F-measure* = 0.65) foram alcançados usando ordem aleatória de *posts*, 100 termos por *post* e 1750 *posts* por usuário. A título de ilustração, os autores discutem uma amostra onde os padrões identificados pelo modelo remetem a situações que podem ser associadas à depressão.

Um estudo recente (MANN; PAES; MATSUSHIMA, 2020) apresenta uma abordagem multimodal para classificar usuários depressivos a partir da mídia social Instagram. O modelo proposto combina duas redes de aprendizado profundo, sendo (1) uma arquitetura ELMo (PETERS et al., 2018a) para processamento das legendas das postagens dos usuários e (2) uma rede ResNet (He et al., 2016), pré-treinada com dados do Instagram, para processamento das imagens. As *features* extraídas por cada rede são combinadas e processadas por uma camada densa de *dropout*, seguida por uma camada linear final que determina a classificação do usuário. Considerando a métrica *F-measure*, os autores reportam que o modelo multimodal alcança 0.79 de performance, sendo superior ao desempenho das redes ELMo (0.75) e ResNet (0.72) quando usadas isoladamente para classificação do usuário. A discussão dos resultados limita-se ao desempenho, sem apresentar uma análise sobre os padrões identificados pelo modelo.

Shen e Rudzicz (2017) abordaram a classificação de *posts* contendo traços de ansiedade em usuários da rede social Reddit. Os autores propõem uma rede neural com duas camadas de profundidade, a qual é combinada a diferentes técnicas de engenharia de *features*. O melhor resultado é alcançado quando o modelo é combinado à abordagem de engenharia de *features* para extração de características léxico-sintáticas usando o dicionário LWIC, somadas ao uso de *n-grams*, ou ao uso de *embeddings* gerados a partir do corpus usando o método Word2Vec. Ambas combinações de *features* apresentam acurácia de 0.98. Os autores apresentam uma análise sobre o conteúdo dos *posts*, demostrando que os termos mais frequentemente encontrados para a classe ansiedade podem ser asso-

ciados com comportamentos típicos de usuários diagnosticados, mas não abordam como relacionar a classificação dos *posts* individuais aos usuários.

Alguns trabalhos aplicam técnicas de aprendizado profundo no contexto de múltiplas desordens, a partir de dados extraídos da rede social Reddit. Gkotsis et al. (2017) exploram duas abordagens de aprendizado profundo, Feed Forward e CNN, para classificação de *posts* em uma entre 11 condições mentais, incluindo ansiedade e depressão. Ambos os modelos de aprendizado profundo incluem uma camada de incorporação de palavras, gerada a partir do próprio corpus. Os autores evoluem os modelos segundo duas abordagens de classificação: (1) binária, na qual *posts* são classificados como saudável ou doente e (2) multi-classe, em que definem a qual desordem um *post* pertence. O estudo não aborda a comorbidade. Os melhores resultados são alcançados com o modelo CNN tanto para a tarefa de classificação binária (acurácia de 91,08%), quanto para a tarefa multi-classe (acurácia média ponderada de 71,37%).

Uma extensão desse trabalho é proposta em (IVE et al., 2018), onde os autores apresentam uma arquitetura RNN (GRU Bi-direcional) para o problema de classificação multi-classe de *posts*. A arquitetura RNN base é formada por uma camada de *embeddings*, uma camada de convolução, uma camada de *max-pooling* e uma camada densa conectada à camada de saída com função sigmoide. Três variações dessa topologia são propostas, sendo as duas primeiras referentes à combinação do modelo RNN com vetores em nível de palavras e sentenças. Em RNN-max é considerado o número máximo, enquanto que RNN-av utiliza o número médio desses vetores. A terceira variação é dada pela configuração RNN com mecanismo de atenção (RNN-att). Todos os modelos RNN atingiram melhor desempenho quando comparados ao modelo CNN desenvolvido anteriormente (GKOTSIS et al., 2017). O melhor resultado foi obtido pelo modelo RNN-att, atingindo *F-measure* ponderada de 76%. Uma análise da distribuição de pesos, segundo o mecanismo de atenção, demonstra que as frases ou termos mais fortemente relacionados à classe da desordem recebem valores de pesos maiores associado à desordem do que expressões "neutras". Os autores concluem que a análise sequencial de texto (RNN) resulta em melhor desempenho para a classificação de *posts* relacionados à saúde mental.

Um trabalho recente (MATERO et al., 2019) explora o potencial das *embeddings* BERT para classificação de usuários com risco de suicídio. O problema é tratado segundo uma abordagem de classificação multi-classe, onde o usuário pode ser classificado em um dos quatro níveis de risco (sem risco, baixo, moderado e severo). Os autores propõem uma abordagem de aprendizado profundo RNN de duplo-contexto, onde o modelo trata

posts associados a fóruns de suicídio separadamente dos demais *posts* para cada usuário através de duas arquiteturas GRU com mecanismo de atenção. O modelo foi testado considerando diferentes combinações de *features* de vocabulário aberto (embeddings BERT), inferência de tópicos (LDA), *features* demográficas (idade e gênero), teóricas (expressões emocionais e traços ansiedade, raiva e depressão) e meta-*features* (estatísticas de postagem e outros 39 *features* derivadas da plataforma Reddit). Os autores concluem que o modelo de contexto duplo produziu ganhos significativos para a tarefa de classificação do risco de suicídio. Embora as dimensões teóricas da linguagem demonstrem estados esperados (linguagem mais depressiva e ansiosa correlacionada com maior risco de suicídio), o modelo atingiu melhor resultado usando as dimensões teóricas combinadas com vocabulário aberto BERT.

Poucos trabalhos abordam a comorbidade entre os transtornos mentais em usuários. Benton, Mitchell e Hovy (2017) propõem um modelo de aprendizado multirrótulo para detectar usuários em risco de suicídio e outras sete condições de saúde mental (isoladas ou em comorbidade), entre elas depressão e ansiedade. O conjunto de dados é composto da comunicação de usuários da rede social Twitter e possui alta taxa de comorbidade entre as desordens. Considerando os 5000 termos mais frequentes, modelam os dados de entrada como *n-grams*, com *n* variando de 1 a 5. Desenvolveram dois modelos *Feed-forward Multilayer Perceptrons* segundo a abordagem de tarefa única (STL) e de multi-tarefa (MTL). Os autores concluem que incluir o gênero do usuário como recurso auxiliar resulta em ganho de performance para muitos transtornos, entre eles a ansiedade. Com exceção da esquizofrenia, o modelo MTL resultou em melhor performance de classificação para as demais desordens isoladas ou em comorbidade. Por fim, treinar o modelo MTL com todas as desordens, mostrou-se benéfico para predizer cada desordem isoladamente. Comparado ao modelo STL para ansiedade, relatam que o modelo MTL proposto reduziu em 11.9% a taxa de erro para classificar ansiedade.

Cohan et al. (2018) exploram a identificação de usuários Reddit em termos de nove condições mentais e suas comorbidades, incluindo ansiedade e depressão. O conjunto de dados é o SMHD utilizado neste trabalho, discutido com mais detalhes no Capítulo 4. Os autores testam tanto as técnicas de aprendizado raso (XBoost, SVM, regressão logística e *FastText*), quanto profundo (CNN). Eles desenvolvem classificadores para cada condição individual (classificação binária de rótulo único), bem como sua comorbidade (classificação multirrótulo). Os resultados foram insatisfatórios em todos os experimentos. A *F-measure* mais alta foi alcançada usando *FastText* (0,54 para classificadores binários de

ansiedade e depressão, 0,27 para o classificador multirrótulo), seguido de CNN (0,47 e 0,50 para os classificadores binários de ansiedade e depressão, respectivamente, e 0,26 para o classificador multirrótulo). O trabalho também buscou identificar a existência de padrões linguísticos que diferenciam os transtornos mentais estudados usando o LIWC, mas nenhum padrão foi encontrado. O trabalho destacou que entre as condições de comorbidade, a co-ocorrência da ansiedade com a depressão é prevalente.

Em resumo, o problema de classificar automaticamente os transtornos mentais é muito complexo e a comorbidade entre eles dificulta ainda mais a identificação dos sintomas que podem influenciar a linha adequada de tratamento. Os trabalhos relacionados contribuíram principalmente com modelos de classificação relacionados a transtornos específicos, quer na forma de classificadores binários ou multiclasse. A Tabela 3.2 sintetiza as principais características dos trabalhos descritos nesta seção.

Tabela 3.2: Identificação de Desordens: Abordagem Supervisionada de Aprendizado Profundo.

Trabalho	**Abordagem de Classificação**	**Transtorno(s)**	**Objetivo (Classificação)**	**Fonte de Dados**	***Insights* sobre Transtorno**	**Principais *Features***	**Modelo**
Yates, Cohan e Goharian (2017)	binário, rótulo único	Depressão e Risco de Auto-mutilação	Usuários	Reddit e ReachOut	Sim, relação de termos e sintomas com base na análise de uma amostra	Variação na ordem dos posts e tamanho de cada mensagem	CNN
Mann, Paes e Matsushima (2020)	binário, rótulo único	Depressão	Usuários	Instagram	Não	Uso de features extraídas de imagem e texto.	Modelo multimodal (ELMo + ResNet)
Shen e Rudzicz (2017)	binário, rótulo único	Ansiedade	Posts	Reddit e Twitter (baseline)	Sim, relação entre termos mais frequentes e sintomas.	Textual (LIWC), Modelagem n-gramas Seleção de Tópicos (LDA) Word embeddings (CBOW) Doc2Vec (PV-DM)	Neural Network
Gkotsis et al. (2017)	multi-classe, rótulo único	Ansiedade, Depressão e outras 9 condições mentais	Posts	Reddit	Não	Word Embeddings	Feed Forward NN + CNN
Ive et al. (2018)	multi-classe, rótulo único	Ansiedade, Depressão e outras 9 condições mentais	Posts	Reddit	Sim, relação entre os pesos produzidos pelo mecanismo de atenção do modelo e sintomas	Word Embeddings	RNN (GRU Bidirecional + mecanismo de atenção)
Matero et al. (2019)	multi-classe, rótulo único	Risco de Suicídio	Usuários	Reddit	Sim, relação de termos e níveis de risco ao suicídio.	Open Vocabulary (embeddings BERT) Seleção de Tópicos (LDA) Recursos Demográficos (idade e gênero) Léxico de Domínio Traços de personalidade *Big-Five* Meta-features	Regressão Logistica e RNN LSTM
Benton, Mitchell e Hovy (2017)	multirrótulo	Ansiedade, Depressão, Comorbidades e outras 6 condições mentais	Usuários	Twitter	Não	Inferência de Gênero N-gramas (N variando de 1 a 5)	Feed-forward Multilayer Percepton
Cohan et al. (2018)	multirrótulo	Ansiedade, Depressão, Comorbidades e outras 7 condições mentais	Usuários	Reddit	Sim, apresenta análises estatísticas para cada desordem.	N-gramas	FastText (*shallow learning*) e CNN
Este Trabalho	**multirrótulo**	**Ansiedade, Depressão e sua Comorbidade**	**Usuários**	**Reddit**	**Sim, relação de *features* influentes com sintomas das doenças.**	**Word Embeddings**	**Ensemble Stacking (CNN, LSTM, Hybrid)**

3.5 Considerações finais

Os trabalhos relacionados fornecem importantes contribuições no que se refere ao desenvolvimento de modelos computacionais dedicados à identificação de transtornos mentais a partir de conteúdo extraído das redes sociais. Os padrões encontrados elucidam ou ratificam comportamentos já identificados nessa área, mas normalmente observados considerando um baixo número de participantes. Contudo, esses estudos apresentam algumas limitações.

Em relação aos trabalhos que abordam a classificação de transtornos mentais com base em técnica de aprendizado raso, verifica-se um grande esforço na etapa de engenharia de *features*, que envolve a definição de recursos extraídos conforme a rede social e transtorno estudado. Como resultado, tipicamente os modelos propostos não apresentam bom desempenho se aplicado a outra rede social ou transtorno diferente.

No que tange o uso de aprendizado profundo para essa tarefa de classificação, os trabalhos relacionados superam essa limitação, uma vez que a abstração de padrões complexos é contemplada em grande parte pelo próprio modelo de aprendizado, dispensando a etapa de engenharia de *features*. Entretanto, dos trabalhos existentes que aplicam essa técnica de aprendizagem para identificação dos transtornos de ansiedade, nenhum aborda a classificação de usuários, apenas de *posts* (SHEN; RUDZICZ, 2017; GKOTSIS et al., 2017; IVE et al., 2018).

Raros trabalhos dedicam-se ao estudo de múltiplos transtornos, incluindo ansiedade e condições de comorbidade (BENTON; MITCHELL; HOVY, 2017; COHAN et al., 2018). Esses trabalhos focaram na classificação de usuários, considerando uma abordagem de aprendizado profundo multi-tarefa, sendo que os modelos propostos não atingiram resultados satisfatórios. Ainda, os *insigths* fornecidos não contemplaram a exploração dos padrões presentes na condição de comorbidade.

O presente trabalho visa contribuir para a redução desta lacuna ao abordar a classificação automática de depressão, ansiedade e sua condição de comorbidade para obter *insights* sobre os padrões comuns e diferenciadores que podem ser derivados da interação social textual. As características marcantes do modelo proposto são o uso de uma abordagem de aprendizado profundo para o desenvolvimento de classificadores usando grandes volumes de dados extraídos de rede social (COHAN et al., 2018), e de técnicas de *ensemble* para superar as dificuldades de lidar com um problema de classificação multirrótulo em um contexto envolvendo a condição de comorbidade.

4 CONJUNTO DE DADOS PARA CLASSIFICAÇÃO DE TRANSTORNOS MENTAIS

Este capítulo é dedicado à apresentação dos conjuntos de dados utilizados para o treinamento do modelo de classificação proposto neste trabalho. Inicialmente, uma visão geral contextualiza a importância do conjunto de dados no desenvolvimento de modelos de aprendizado de máquina. Em seguida, apresenta-se a base de dados SMHD juntamente com a metodologia adotada para derivação dos conjuntos de dados usados neste trabalho.

4.1 Processo de Criação do Conjunto de Dados SMHD

A técnica usada para rotular o conjunto de dados SMHD foi originalmente proposta por Yates, Cohan e Goharian (2017) no contexto de depressão. Posteriormente, a mesma técnica foi estendida para considerar outras nove condições mentais (COHAN et al., 2018), gerando o conjunto de dados SMHD. O método propõe o uso de padrões de alta precisão para identificar usuários que alegam terem sido diagnosticados com uma condição de saúde mental (denominados usuários diagnosticados) e usam critérios de exclusão para combiná-los com usuários de controle. O método foi projetado para evitar viés entre os grupos controle e diagnosticado, de modo que a tarefa de classificação não seja artificialmente fácil devido à presença de expressões óbvias.

O método proposto para a formação do conjunto de dados segue as etapas detalhadas na Seção 3.2. Inicialmente, uma busca por usuários diagnosticados em fóruns (*subreddits*) de saúde mental do Reddit é realizada, usando como filtro de pesquisa expressões de auto-diagnóstico (por exemplo, "eu fui diagnosticado"). Em seguida, uma inspeção manual é realizada com a finalidade de remover falsos positivos, isto é, casos em que a auto-declaração do usuário reflete um sentido de dúvida, brincadeira ou mesmo afirma não ser diagnosticado (por exemplo, "eu nunca fui diagnosticado clinicamente"). Posteriormente, os candidatos a usuários de controle para cada transtorno são coletados de acordo com os seguintes critérios (1) nunca ter publicado em um *subreddit* de saúde mental e (2) nunca ter utilizado termos relacionados à saúde mental. Os autores selecionaram os usuários de controle com a maior probabilidade de postar nos mesmos *subreddits* que um usuário diagnosticado, resultando em uma proporção aproximada de 1:9. Após coletar o histórico de postagens para ambos os tipos de usuários, o processo é finalizado com

uma etapa de limpeza de dados. Nessa etapa, foram removidas todas postagens pertencentes aos *subreddits* de saúde mental. Também foram removidas mensagens pertencentes a outros fóruns, mas que incluíam linguagem relacionada à saúde mental. Por exemplo, mensagens contendo os termos "deprimido", "diagnóstico"ou "doença mental"foram excluídas, garantindo assim que usuários diagnosticados não sejam facilmente identificados pela presença de termos obviamente relacionados ao transtorno mental.

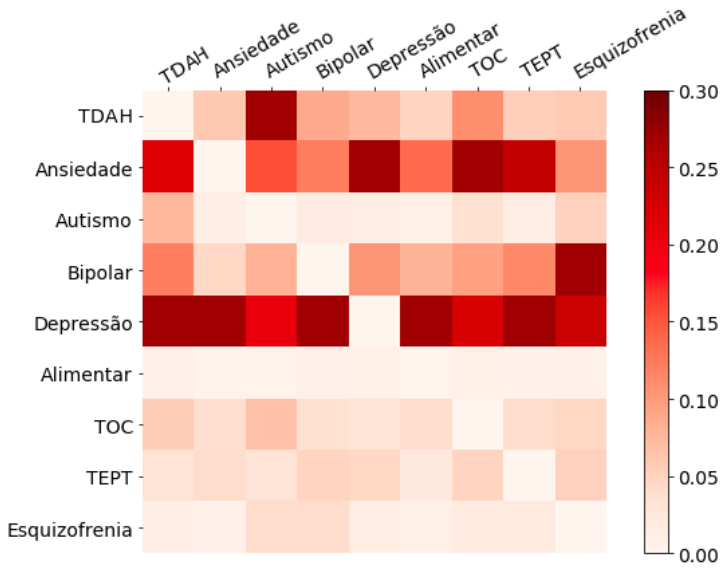
A validação do processo de rotulagem do conjunto de dados foi realizada de modo totalmente manual em (YATES; COHAN; GOHARIAN, 2017). A classe de usuários diagnosticados foi definida por votação majoritária, com base na classificação arbitrada por três anotadores não especializados em saúde mental. Já no conjunto de dados SMHD (COHAN et al., 2018), o reconhecimento de sentenças de auto-declaração foi realizado por uma abordagem de rotulagem automática de dados que combina dois métodos para mapeamentos de sinônimos, a saber, *MedSym* (YATES; GOHARIAN, 2013) e o *Behavioral* (YOM-TOV; GABRILOVICH, 2013). Esses métodos permitem mapear sinônimos para expressões comumente usadas por leigos para descrever e pesquisar sua condição médica na Internet. Por fim, uma etapa de inspeção manual foi realizada, considerando uma subamostragem do conjunto final SMHD. Essa inspeção revelou uma precisão de 96% no método de rotulagem automática para todas as condições, exceto para ansiedade (90%). Os autores apontaram que muitos falsos positivos foram causados por termos muito próximos para definir um diagnóstico entre duas ou mais condições mentais.

4.2 Estatísticas do Conjunto de Dados SMHD

Os transtornos mentais presentes no conjunto de dados SMHD foram selecionados com base nas seis condições consideradas mais frequentes segundo o manual DSM-5, sendo elas a Esquizofrenia, Bipolaridade, Depressão, Ansiedade, Transtorno Obsessivo Compulsivo (TOC) e Desordens Alimentares. Foram ainda acrescidos de outras três condições menos frequentes: TEPT, Transtorno de Déficit de Atenção/Hiperatividade (TDAH) e Autismo.

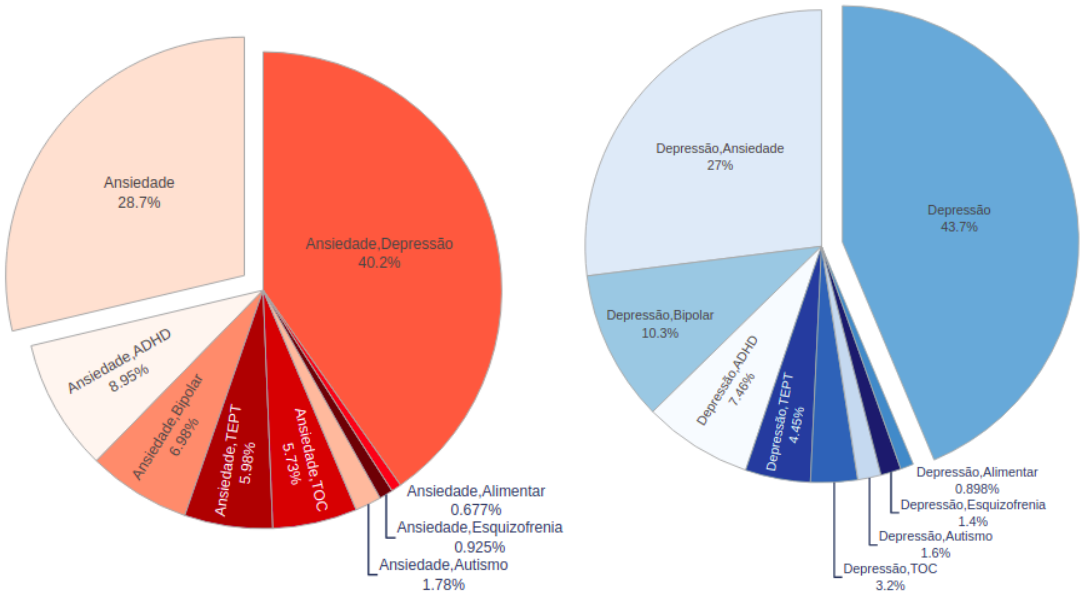
Ao todo, o conjunto de dados SMHD é composto por 20.406 usuários diagnosticados e 335.952 controles correspondentes, considerando um período de coleta de postagens que compreende de janeiro/2006 a dezembro/2017. Do total de usuários diagnosticados, 8,5% sofrem somente de ansiedade, e 19,3% somente de depressão. Ainda, 26,7% dos usuários correspondem a condições de comorbidade, as quais são definidas no conjunto

Figura 4.1: Comorbidade: Matriz de co-ocorrência entre os transtornos mentais.



Fonte: Cohan et al. (2018)

Figura 4.2: Posicionamento dos Transtornos Alvo em Relação às Comorbidades.



(a) Distribuição da Ansiedade e Comorbidades (b) Distribuição da Depressão e Comorbidades

de dados SMHD pela presença de dois ou mais transtornos mentais (e.g. ansiedade e depressão; depressão, TOC e ansiedade). Entre as condições de comorbidade, verifica-se que a depressão, seguida da ansiedade são os transtornos que mais co-ocorrem com outros transtornos mentais, como mostra a Figura 4.1. Ainda, quase 30% dos usuários com depressão, TOC ou TEPT também sofrem de ansiedade (COHAN et al., 2018).

A Figura 4.2 (a) apresenta a distribuição do total de usuários diagnosticados com transtornos de ansiedade, isoladamente ou em comorbidade com outras doenças. Verifica-se que os quatro transtornos mentais com maior taxa de comorbidade com a ansiedade são

a depressão, seguido do TDAH, bipolaridade e TEPT. Por sua vez, a depressão também apresenta alta taxa de comorbidade para os mesmos transtornos, apresentando uma ordem de prevalência diferente: ansiedade, bipolaridade, TDAH e TEPT, como mostra a Figura 4.2 (b). Ainda, é possível constatar que a manifestação do transtorno de ansiedade isoladamente é menos frequente do que em condição de comorbidade com outros transtornos. Já na depressão, observa-se que sua manifestação isolada é mais frequente. A relação quantitativa observada nos dados do conjunto SMHD para esses transtornos ratifica os achados de estudos relacionados (HIRSCHFELD, 2001; TILLER, 2013), os quais afirmam que a manifestação de ansiedade é observada em 85% dos pacientes depressivos, assim como 90% dos pacientes ansiosos apresentam comorbidade com a depressão.

O conjunto de dados SMHD foi disponibilizado via acordo de uso de dados firmado com a Universidade de Georgetown, o qual foi respeitado no presente trabalho. Os curadores dividiram esse conjunto de dados em três subconjuntos dedicados ao treinamento, validação e teste de modelos preditivos, sugerindo que essa divisão fosse mantida para estudos posteriores. Cada subconjunto contém uma distribuição de dados semelhante em termos de total de usuários e postagens por usuários. Cabe destacar que cada usuário, juntamente com seu conjunto completo de postagens, está presente somente em um dos subconjunto de dados gerados.

Para atender o objetivo do presente trabalho, optou-se por gerar subconjuntos de dados que contenham apenas as três condições de saúde mental foco desse estudo acrescidas de seus usuários de controle. A seção a seguir descreve a formação desses subconjuntos.

4.3 Conjunto de Dados para Treinamento do Modelo Proposto

Para investigar cada transtorno mental individualmente e sua comorbidade, foram derivados sete conjunto de dados do SMHD, sumarizados na Tabela 4.3. Neste trabalho, denominaremos de condição mental ou condição alvo, um transtorno mental específico (Ansiedade ou Depressão), ou a comorbidade desses transtornos (Ansiedade e Depressão). Os seis primeiros conjuntos de dados foram criados com intuito de atender a premissa de explorar o treinamento dos classificadores binários (rótulo único), que compõem o nível inferior do modelo *DAC Stacking*, quanto a função que exercem:

- *Classificadores Especializados em identificar cada Condição Mental*: estes mode-

los são usados para distinguir usuários diagnosticados de saudáveis (controle). Para treiná-los, foram criados três conjuntos de dados de rótulo único, um para cada condição, juntamente com os respectivos usuários de controle. As condições mentais são Ansiedade (A), Depressão (D) ou Comorbidade (AD);

- *Classificadores Diferenciadores entre Condições Mentais*: estes classificadores visam distinguir entre condições mentais diagnosticadas. Para treiná-los, foram criados três conjuntos de dados de rótulo único, um para cada par de condições mentais: Ansiedade ou Depressão (A-D), Ansiedade ou Comorbidade (A-AD) e Depressão ou Comorbidade (D-AD). Esses conjuntos de dados não incluem usuários de controle.

Por fim, o sétimo conjunto de dados (A-D-AD) foi preparado para o treinamento do modelo *DAC Stacking*, o qual é treinado segundo a abordagem de classificação multirrótulo, sendo formado por amostras de usuários diagnosticado com um ou ambos transtornos, junto com os respectivos usuários de controle.

Os conjuntos de dados da Tabela 4.3 foram gerados de modo balanceado visando evitar o viés durante o processo de aprendizagem dos classificadores. Para tanto, foram considerados os seguintes critérios de seleção de usuários: (1) total de *posts* por classe de usuário (condição alvo ou controle); e (2) usuários com o histórico de postagem de pelo menos 50 *posts*.

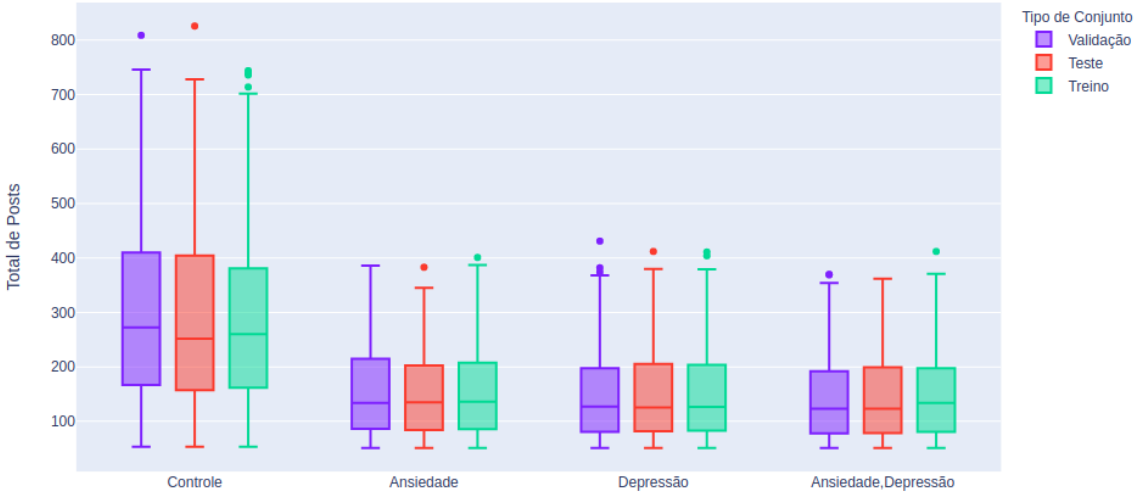
Cada conjunto de dados gerado manteve a divisão original de amostras, segundo o

Figura 4.3: Conjunto de dados derivados de SMHD para os experimentos.

Conjunto de Dados	**Tipo de Rótulo**	**Classe**	**Total de Amostras**					
			Treinamento		**Validação**		**Teste**	
			Usuários	***Posts***	**Usuários**	***Posts***	**Usuários**	***Posts***
SMHD A	rótulo único	Ansiedade	520	81.541	520	80.596	520	78.193
		Controle	520	153.965	520	154.790	520	149.609
SMHD D	rótulo único	Depressão	1.073	160.935	1.077	157.904	1.080	155.432
		Controle	1.080	310.285	1.080	313.990	1.080	307.984
SMHD AD	rótulo único	Ansiedade, Depressão	440	65.307	440	62.698	440	63.051
		Controle	440	133.964	440	130.755	440	126.173
SMHD A-D	rótulo único	Ansiedade	520	81.541	520	80.596	520	78.193
		Depressão	520	78.757	520	76.395	520	75.062
SMHD A-AD	rótulo único	Ansiedade	440	70.316	440	68.781	440	67.167
		Ansiedade, Depressão	440	65.307	440	62.698	440	63.051
SMHD D-AD	rótulo único	Depressão	440	67.547	440	64.948	440	64.746
		Ansiedade, Depressão	440	65.307	440	62.698	440	63.051
SMHD A-D-AD	multi-rótulo	Ansiedade	440	67.982	440	68.418	440	65.970
		Depressão	440	65.549	440	64.900	440	65.262
		Ansiedade, Depressão	440	65.307	440	62.698	440	63.051
		Controle	880	252.500	880	257.172	880	254.502

objetivo de treinamento, validação e teste. A importância de manter essa divisão está em garantir que cada classe de usuário apresente um perfil semelhante em cada subconjunto, minimizando problemas de viés durante a etapa de treinamento e validação dos modelos. Uma inspeção foi realizada para verificar o perfil do usuário em cada subconjunto quanto à distribuição de *posts* por usuário, conforme mostra a Figura 4.4. Verifica-se que a distribuição dos usuários em cada classe é mantida semelhante nos subconjuntos de treino, validação e teste. Ainda, a distribuição de *posts* revela que usuários saudáveis, em média, postam mais que usuários diagnosticados. Entre os usuários diagnosticados, observa-se um comportamento semelhante para a distribuição do total de postagens.

Figura 4.4: Distribuição de *Posts* por Usuário e Tipo de Conjunto de Dados



4.4 Considerações finais

Este capítulo descreveu os processos de formação dos conjuntos de dados SMHD e os derivados a partir dele, usados neste trabalho com a finalidade de treinar os modelos dedicados à classificação dos transtornos de ansiedade, depressão e sua comorbidade. Os critérios de seleção para formação desses conjuntos derivados mantém a distribuição de usuários por classe, segundo os subconjuntos de treinamento, validação e teste, conforme sugerido por Cohan et al. (2018).

O uso do conjunto de dados SMHD para este estudo apresenta algumas vantagens. Primeiramente, a formação desse conjunto de dados contempla o uso de um método de rotulagem automática altamente preciso, projetado para evitar viés, o que torna seu uso adequado ao treinamento de modelos de aprendizado de máquina. Outra vantagem é a

sua formação a partir da extração de conteúdos dos usuários da rede social Reddit, a qual possibilita que os mesmos criem fóruns, *subreddits*, de tema livre e sem limitação de palavras para postagens. Essas características estimulam os usuários a exporem seus sentimentos e opiniões de modo mais detalhado, o que beneficia este estudo.

Por fim, é importante destacar que os conjuntos de dados usados não representam uma totalidade em termos de comportamento para os usuários diagnosticados. Entretanto, devido à abrangência e adesão mundial da rede social Reddit, pode-se considerar que os conjuntos de dados representam uma subpopulação variada para o estudo de transtornos mentais.

5 UMA ABORDAGEM DE COMITÊ PARA A CLASSIFICAÇÃO DE TRANSTORNOS MENTAIS

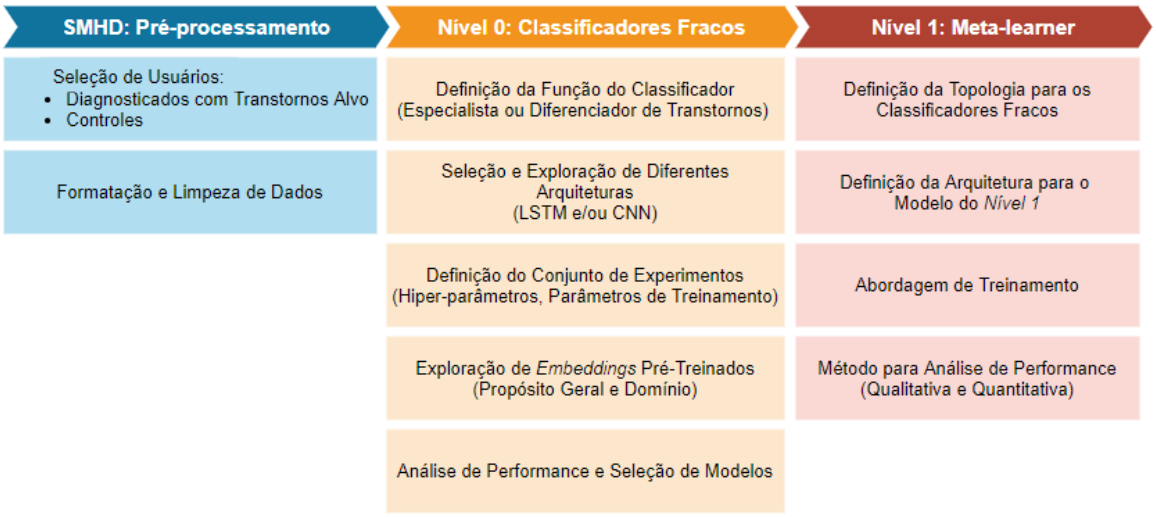
Este capítulo apresenta a metologia empregada para o desenvolvimento do modelo proposto, *DAC Stacking*, para a classificação de usuários diagnosticados com ansiedade, depressão ou comorbidade desses transtornos a partir da análise de suas publicações na rede social Reddit. Inicialmente, apresenta-se uma visão geral sobre os principais aspectos envolvidos no desenvolvimento do modelo proposto. As seções seguintes, dedicam-se a descrição detalhada de cada etapa desse processo.

5.1 Visão Geral da Proposta

Como visto no Capítulo 3, poucos trabalhos dedicam-se à classificação automática dos transtornos de ansiedade a partir de conteúdo de mídias sociais. A comorbidade desse transtorno com a depressão é ainda menos explorada. Nesse trabalho, os transtornos de ansiedade e depressão são abordados considerando a perspectiva de solução para problemas de classificação multirrótulo. Nessa abordagem, um usuário pode ser identificado com um ou mais transtornos, contemplando assim a condição de comorbidade. A classificação multirrótulo foi escolhida devido à vantagem de tornar a solução proposta extensível a outras condições mentais e possíveis comorbidades. O modelo proposto é treinado para identificar os traços dos transtornos mentais com base no histórico de mensagens do usuário da rede social Reddit, mas pode ser adaptado para outras estruturas de comunicação que seguem um padrão análogo.

O emprego da abordagem de comitê para o desenvolvimento do modelo proposto visa cobrir as dificuldades de lidar com um problema de classificação multirrótulo envolvido em um cenário de comorbidade, onde a distinção de padrões é uma tarefa difícil. Nesse cenário, combinar classificadores especializados em cada transtorno pode ser mais efetivo para a identificação da condição de comorbidade, do que o desenvolvimento de um único modelo. Desse modo, *DAC Stacking* segue uma abordagem *ensemble*, onde o nível inferior (*Nível 0*) é composto de classificadores binários de rótulo único (*single-label*) que preveem probabilidades para a classe do usuário, segundo as opções controle (saudável) ou diagnosticado com uma condição-alvo específica (ansiedade, depressão ou comorbidade). No nível superior (*Nível 1*), essas previsões individuais são consolidadas

Figura 5.1: Metodologia para desenvolvimento do modelo *DAC Stacking*.



usando uma rede neural densa, que lida com o problema de classificar o usuário segundo as atribuições de rótulo de controle ou diagnosticado com um ou ambos transtornos mentais.

A Figura 5.1 sintetiza as principais etapas do processo de desenvolvimento do *DAC Stacking*, detalhadas neste capítulo. Entre os principais desafios envolvidos na formação desse modelo, destacam-se:

a) *Composição do Nível 0*: abrange as definições referentes às arquiteturas usadas para o desenvolvimento dos classificadores fracos, bem como as abordagens empregadas para gerar variabilidade entre esses modelos. Para atender a primeira questão, foram exploradas as arquiteturas LSTM, CNN e modelos híbridos através de um extenso conjunto de experimentos envolvendo variações dos hiperparâmetros e parâmetros de treinamento para cada arquitetura. Para gerar variabilidade entre os modelos fracos, foram adotadas as seguintes estratégias: (1) treinar diferentes classificadores binários variando sua função quanto a ser *especialista* em cada condição-alvo (ansiedade, depressão e comorbidade), ou *diferenciador* entre essas condições; e (2) o uso de diferentes *embeddings* pré-treinados de propósito geral e de domínio (gerados a partir do próprio *corpus*);

b) *Composição do Nível 1*: engloba as definições acerca da arquitetura do *ensemble* e abordagem de treinamento usadas para a formação do modelo *meta-learner*, bem como da topologia do *Nível 0* em termos dos classificadores fracos;

c) *Análise de Performance*: compreende a definição dos métodos para a avaliação

quantitativa e qualitativa do modelo proposto. Para análise quantitativa, adotou-se um conjunto de métricas específicas para avaliação de desempenho tanto dos classificadores binários de rótulo único, quanto do modelo final multirrótulo, detalhado na Seção 2.6. Para a avaliação qualitativa, realizou-se a interpretação dos modelos gerados, relacionando as características mais influentes para classificação com os sintomas conhecidos em cada transtorno, com uso da abordagem SHAP.

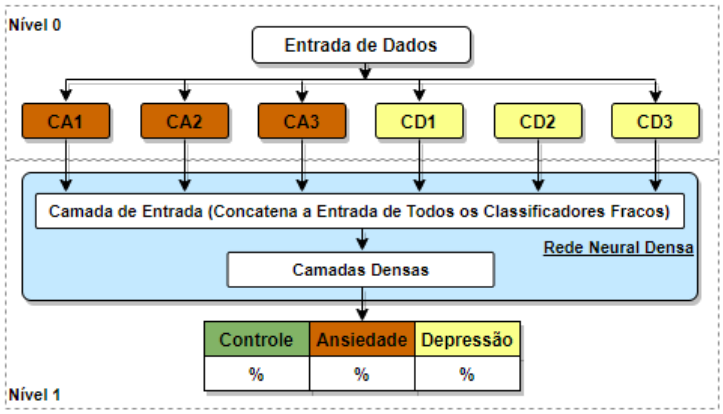
5.2 *DAC Stacking*: Arquitetura

Como já mencionado, o modelo *DAC Stacking* foi desenvolvido sob uma abordagem de empilhamento considerando dois níveis. No *Nível 0*, os classificadores fracos são modelos binários de rótulo único projetados para exercer diferentes funções quanto à tarefa de classificação. No *Nível 1*, uma rede neural densa é responsável por aprender a ponderar as decisões individuais de cada classificador fraco e emitir a predição final no formato multirrótulo. Para determinar a presença de cada rótulo, convencionou-se a probabilidade final de ocorrência da classe ser maior que 50%. Essa estrutura é mantida a mesma em todas as variações propostas para o modelo *DAC Stacking*.

Para desenvolver os classificadores fracos com funções diferentes, foram gerados conjuntos de dados específicos para o treinamento desses classificadores, já detalhados na Seção 4.3 (Tabela 4.3). Na função de especialista, os classificadores se dedicam à identificação de cada transtorno, sendo treinados usando os conjuntos de dados SMHD A (Ansiedade), D (Depressão) e AD (Comorbidade). Já os classificadores diferenciadores têm como função distinguir padrões entre os transtornos e são treinados com os conjuntos de dados SMHD A-D (Diferenciadores entre Ansiedade e Depressão), A-AD (Diferenciadores entre Ansiedade e Comorbidade) e D-AD (Diferenciadores entre Depressão e Comorbidade).

A Figura 5.2 ilustra a topologia base do modelo *DAC Stacking*, a qual é definida por um conjunto de seis classificadores fracos especialistas, onde três são dedicados à identificação do transtorno de ansiedade (CA_i) e os demais ao transtorno de depressão (CD_i). A partir dessa topologia base, duas variações foram propostas visando explorar a combinação de funções para os classificadores fracos que resulta em maior ganho de performance para reconhecer a condição de comorbidade. Essas variações foram produzidas pelo acréscimo de mais três classificadores fracos à arquitetura base, gerando os modelos:

Figura 5.2: Arquitetura Base *DAC Stacking*.



- *DAC Stacking EC (Especialistas em Comorbidade)*: formado pela inclusão dos classificadores especialistas na condição de comorbidade, identificados na Figura 5.3 pelo conjunto de classificadores CAD_i;

- *DAC Stacking DT (Diferenciadores de Transtornos)*: gerado pela inclusão dos classificadores fracos diferenciadores entre duas condições mentais, como mostra a Figura 5.4. Nela, os classificadores diferenciadores estão identificados pelas siglas CA-D (Ansiedade e Depressão), CA-AD (Ansiedade e Comorbidade) e CD-AD (Depressão e Comorbidade).

Além das funções, os classificadores fracos foram combinados considerando as diferentes arquiteturas de redes profundas usadas para a formação desses modelos. Assim, exitem *DAC Stackings* homogêneos, onde todos os classificadores fracos seguem uma única arquitetura, ou heterogêneos, que contém classificadores fracos de diferentes arquiteturas.

Figura 5.3: Arquitetura Base *DAC Stacking EC*.

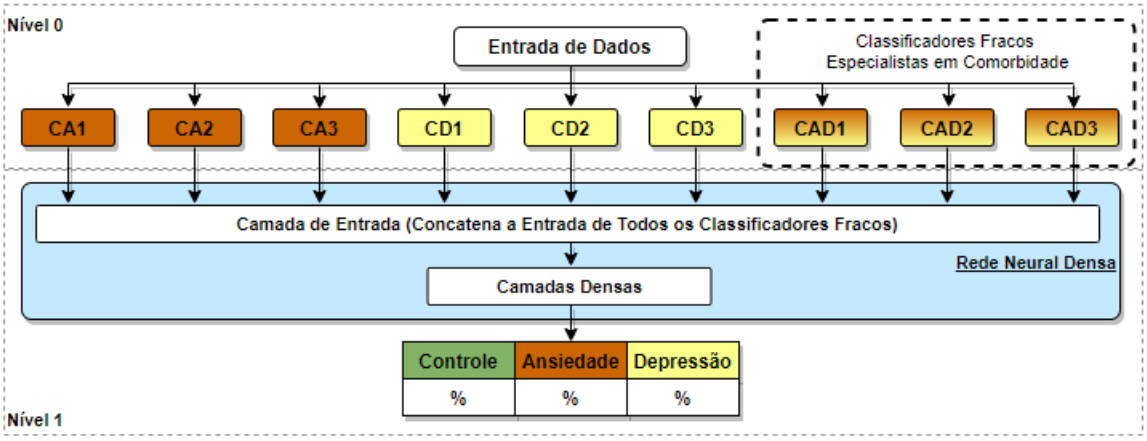
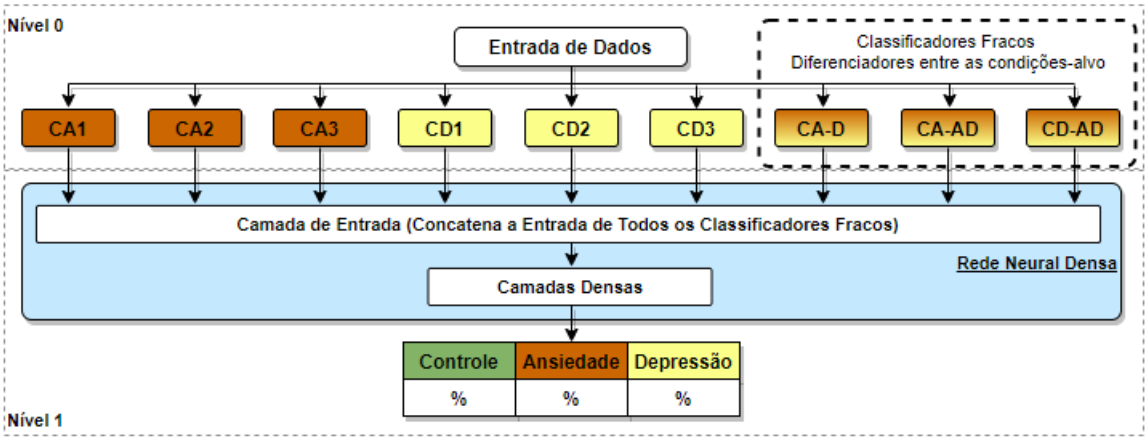


Figura 5.4: Arquitetura Base *DAC Stacking DT*.



O modelo *DAC Stacking* e suas variações são resultado de um extenso conjunto de experimentos, os quais foram desenvolvidos usando o pacote científico Python 3.6, Keras 2.2.5[1] com *backend* TensorFlow 1.14.0[2]. A implementação desses experimentos, bem como o projeto completo para gerar o modelo proposto está disponível em repositório público no GitHub[3]. As escolhas referentes ao desenvolvimento de classificadores fracos, à definição da rede neural que compõe a camada *meta-learner* e métodos de avaliação do modelo são descritas em detalhes nas próximas seções.

5.3 Pré-processamento dos Dados de Entrada

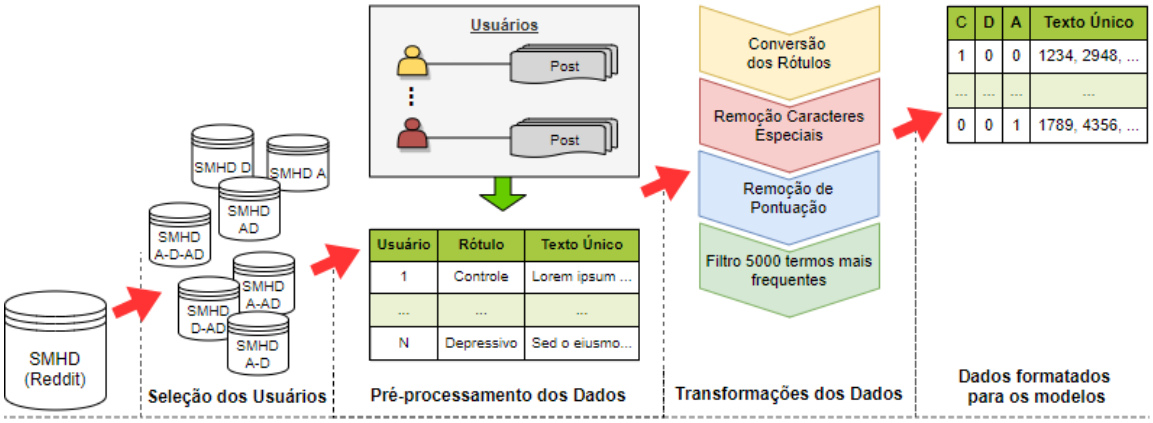
Como resume a Figura 5.5, a etapa de pré-processamento compreendeu preparar os dados para o treinamento dos modelos de classificação quanto à seleção dos usuários de interesse, ao formato de entrada e à limpeza dos dados.

A etapa inicial de seleção de amostras contempla gerar subconjuntos de dados, formados por usuários diagnosticados com os transtornos alvo deste estudo e respectivos usuários de controle. Esse processo já foi descrito em detalhes na Seção 4.3, juntamente com os conjuntos de dados resultantes.

Quanto ao formato de entrada para os dados, a definição de como os *posts* dos usuários seriam alimentados na rede neural foi realizada por processo de experimentação. Nesses experimentos, a representação do histórico de postagens do usuário foi explorada segundo dois formatos: (1) *lista de posts* e (2) *texto único*. O segundo formato atingiu os

[1] https://keras.io/
[2] https://www.tensorflow.org/
[3] https://github.com/borbavanessa/dac_stacking

Figura 5.5: Pré-processamento dos Dados: Estrutura Comum



melhores resultados, sendo adotado como padrão para os modelos *DAC Stacking*. Nesse formato, o usuário é representado por sua sequência completa de *posts*, concatenados conforme ordem cronológica, formando um único texto.

O processo de formatação dos dados de entrada é finalizado com o uso da função de *tokenização*[4], a qual contempla as etapas de conversão do texto em *tokens* e limpeza dos dados. Desse modo, o texto concatenado para cada usuário é processado por essa função, a qual foi parametrizada para formatar o texto para letra minúscula, remover caracteres especiais (e.g. pontuações e símbolos) e extrair os 5000 termos mais frequentes. Para o rótulo da classe do usuário, adotou-se o padrão de conversão *one-hot enconding*, no qual cada classe é representada por uma posição fixa em um vetor, cuja as posições com valor 1 indicam a presença do rótulo para a classe.

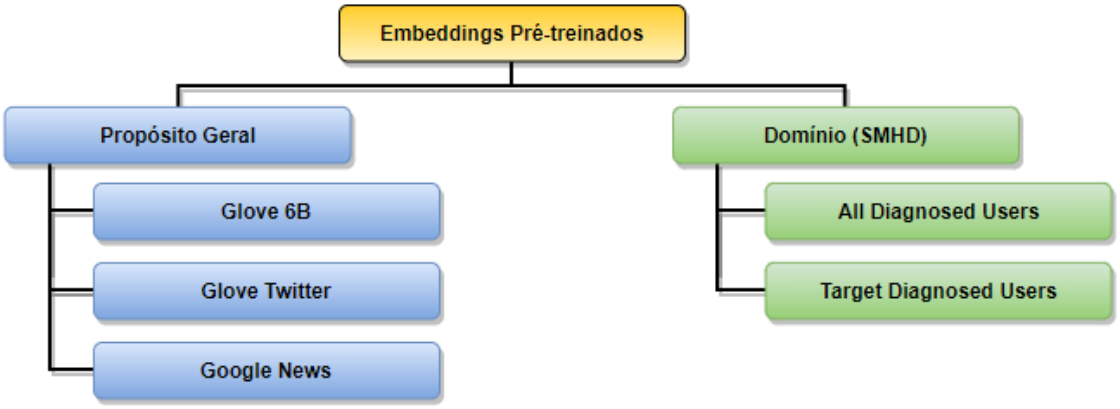
5.4 *Embeddings* Pré-treinados

Com a finalidade de gerar variabilidade entre os classificadores fracos foram experimentadas diferentes configurações para a camada de *embeddings*, conforme detalhado no Apêndice A. A estratégia de combinar classificadores fracos que usam diferentes *embeddings* foi motivada pela premissa de que o uso de *embeddings* gerados a partir de conjuntos de dados diferentes permitiria recuperar contextos complementares para um mesmo termo, auxiliando na tomada de decisão dos modelos para a classificação das condições-alvo. Assim, um segundo conjunto de experimentos foi projetado para avaliar o impacto de variar os *embeddings* pré-treinados na performance dos classificadores fracos, em relação à arquitetura (LSTM e CNN) e condição-alvo (ansiedade, depressão

[4]https://keras.io/api/preprocessing/text/#tokenizer-class

e comorbidade). Esses experimentos foram divididos, segundo o critério fonte de dados usada para a formação dos *embeddings* pré-treinados, em *propósito geral* e *domínio*. A Figura 5.6 sintetiza os *embeddings* explorados.

Figura 5.6: *Embeddings* Pré-treinados: Propósito Geral x Domínio



O primeiro grupo de experimentos contemplou o uso de *embeddings* formados a partir de conteúdos extraídos de serviços *online* e disponibilizados publicamente para uso. Desse modo, foram usados os *embeddings* Glove 6B, Glove Twitter e Google News.

O segundo grupo de experimentos contempla os *embeddings* de domínio, gerados por este trabalho a partir do corpus SMHD (COHAN et al., 2018). Cada *embedding* pré-treinado foi explorado nos modelos considerando tanto a abordagem de aprendizado estático, quanto não-estático.

Para a formação dos *embeddings* de domínio foram empregados os algoritmos Glove e Word2Vec (Skip-gram e CBOW) para a formação de dois conjuntos de *embeddings*. O primeiro, denominado *All Diagnosed Users*, considera o conteúdo dos *posts* de todos os usuários diagnosticados no conjunto de dados SMHD (COHAN et al., 2018). Isso representa considerar o vocabulário empregado por usuários diagnosticados com outros trantornos mentais, incluindo ansiedade e depressão. Na segundo, nomeado *Target*

Tabela 5.1: Parametrização dos algoritmos para formação dos *embeddings* de domínio.

Abordagem	**Fonte de Dados (SMHD)**	**Parâmetro**	**Valor**
Word2Vec	All Diagnosed Users e Target Diagnosed Users	Embedding Size	300
		Window Size	5
		Minimun Count Frequence	5
		Number worders	6
		Algorithm	Skip-gram, CBOW
Glove	All Diagnosed Users e Target Diagnosed Users	Window Size	5
		Number of components	300
		Learning rate	0,25
		Epochs	30
		Number of Threads	6

Diagnosed Users, os *embeddings* foram gerados a partir de *posts* de usuários diagnosticados somente com os transtornos alvo desse estudo. A Tabela 5.1 apresenta os parâmetros definidos em cada algoritmo para a formação dos *embeddings* de domínio, que correspondem à parametrização padrão sugerida pelas bibliotecas Glove e Gensim. Experimentos demonstrando os efeitos destes diferentes *embeddings* no desempenho dos classificadores especialistas são detalhados na Seção 6.3.

5.5 *Nível 0*: Classificadores Fracos

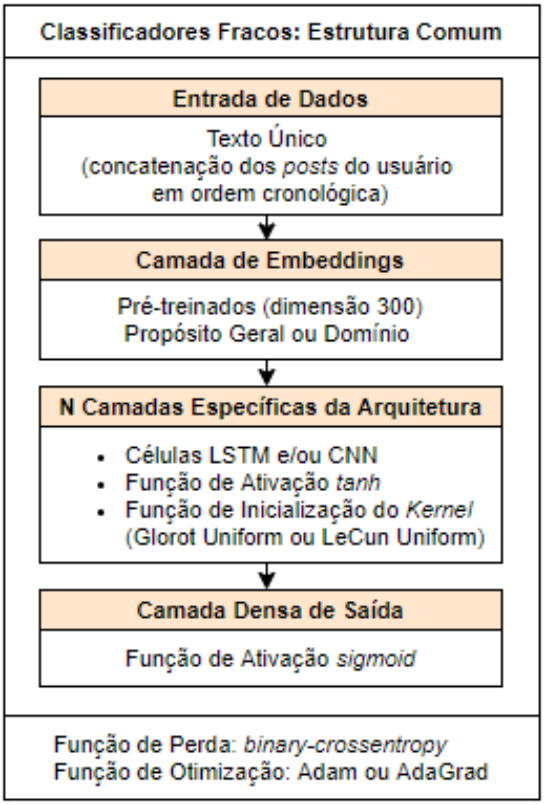
O desenvolvimento dos classificadores fracos tanto *especialistas*, quanto *diferenciadores* de transtornos mentais, compreendeu a exploração de três arquiteturas de aprendizado supervisionado profundo. Essas arquiteturas foram selecionadas por apresentarem diferentes princípios de aprendizagem, sendo elas:

- *LSTM*: escolhida com base na premissa de que a sequência temporal de postagens dos usuários poderia alavancar a descoberta de padrões para as condições alvo;
- *CNN*: devido à suposição de que sua capacidade de identificar padrões locais, considerando diferentes níveis de complexidade, poderia auxiliar na distinção entre as desordens;
- *Modelos Híbridos*: compostos por camadas CNN e LSTM, com o intuito de explorar o potencial das duas arquiteturas, de modo que diferentes padrões locais possam ser explorados ao longo de sequências temporais.

A Figura 5.7 ilustra a topologia genérica dos classificadores fracos, cuja estrutura comum possui os seguintes elementos:

1. *Camada de Entrada*, que define o formato esperado para os dados de entrada, isto é, como o texto do usuário será processado pelo modelo;
2. *Camada de Embeddings*, a qual enriquece com *embeddings* pré-treinados termos de entrada do conteúdo de cada usuário, agregando vetores adicionais para cada termo;
3. *Camadas Específicas à Arquitetura (LSTM, CNN ou ambas)*, compreende a configuração dos hiperparâmetros gerais da rede, bem como os específicos a cada arquitetura;

Figura 5.7: Topologia dos Classificadores Fracos: Estrutura Comum



4. *Camada Final*, configurada para usar a função de ativação *sigmoid* e função de perda *binary crossentropy*, definidas segundo padrão indicado para solução de problemas de classificação binária (CHOLLET, 2017).

Além disso, foram experimentadas duas variações para a função *kernel* de inicialização da rede, Glorot[5] e LeCun[6], as quais diferem em termos da função de distribuição usada para iniciar a matriz de pesos *kernel* usada para a transformação linear das entradas. Uma função de inicialização do *kernel* adequada pode reduzir/evitar problemas relacionados à convergência da rede durante o aprendizado, particularmente relacionados à explosão/decaimento dos gradientes.

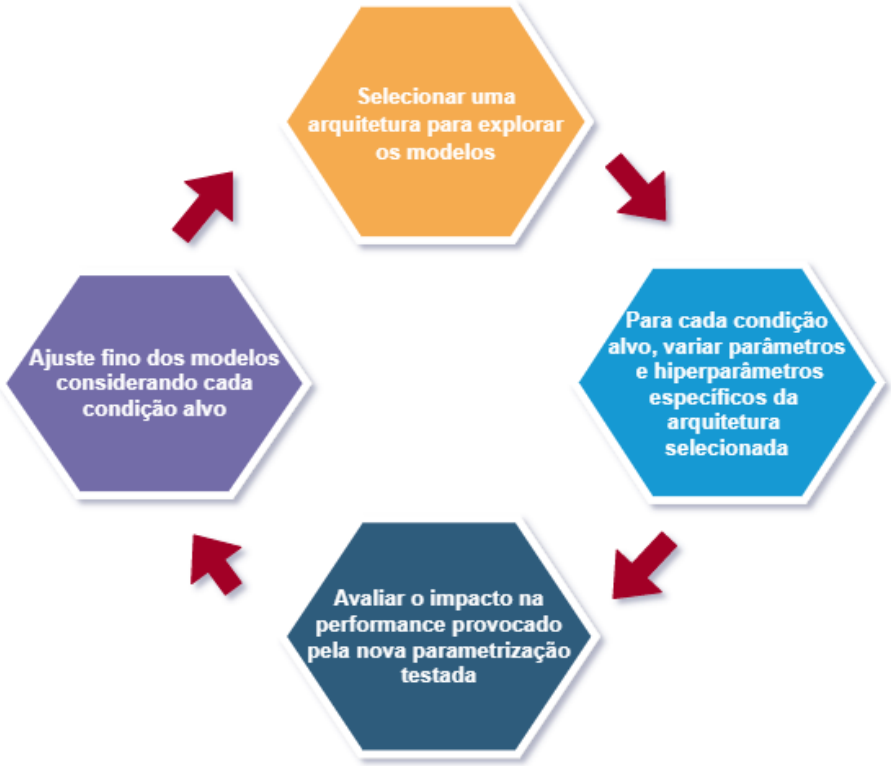
O desenvolvimento dos classificadores fracos envolveu um extenso conjunto de experimentos para definição da parametrização dos modelos que resultassem no melhor desempenho de classificação, conforme função do classificador. O resultado desses experimentos são detalhados na Seção 6.2. O restante dessa seção descreve o processo de formação dos classificadores fracos especialistas e diferenciadores entre condições mentais.

[5]http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf
[6]http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf

5.5.1 Classificadores Especialistas em Condições Mentais

Independente da arquitetura, os classificadores fracos especialistas em transtornos mentais foram desenvolvidos segundo o fluxo de experimentos ilustrado na Figura 5.8. Esse processo cíclico de experimentos foi realizado considerando cada um dos conjuntos de dados SMHD A, D e AD (Tabela 4.3). Os três melhores resultados de cada arquitetura foram selecionados para compor os modelos *DAC Stacking* e suas variações.

Figura 5.8: Principais Etapas para formação do classificadores especialistas.



Os experimentos visaram encontrar uma topologia base mínima para cada classificador especialista e, partir dessa topologia, uma etapa final de experimentos foi empregada para ajuste fino dos modelos em relação a cada transtorno mental. A arquitetura LSTM foi a primeira explorada, onde um conjunto mais extenso de experimentos foi desenvolvido. Os experimentos com as arquiteturas CNN e Híbrida foram construídos com base no conhecimento sobre parametrização de modelos neurais adquirido ao longo do conjunto de experimentos, e não foram tão exaustivos.

5.5.1.1 Arquitetura LSTM

O presente trabalho iniciou pela exploração da arquitetura LSTM para função especialista em cada condição alvo, a qual envolveu sucessivos conjuntos de experimentos

(SOUZA; NOBRE; BECKER, 2020). Como a arquitetura LSTM apresenta diferentes hiperparâmetros e seu treinamento requer um grande tempo de execução, adotou-se a estratégia de definir um subconjunto de parâmetros, os quais são considerados os mais críticos para variação de performance dos modelos LSTM (CHOLLET, 2017).

A Tabela 5.2 apresenta a arquitetura base LSTM resultante desse processo de experimentação. Essa topologia foi usada como base para os experimentos que exploram, respectivamente, o impacto de variar a função *kernel* para inicialização dos pesos da rede e as contribuições oriundas de variar os *embeddings*, quanto a pré-treinados de domínio. De modo geral, os conjuntos de experimentos incluíram as seguintes variações:

a) *Embeddings*: diferentes configurações para a camada de *embeddings* foram explorados, conforme detalhado na Seção 5.4. De modo geral, os modelos formados com uso de *embeddings* pré-treinados de propósito geral apresentaram melhor desempenho. Como uma estratégia para aumentar a variabilidade da solução final, foram selecionados os melhores resultados considerando os *embeddings* pré-treinados de propósito geral e de domínio;

b) *Funções de Ativação*: foram testadas as funções de ativação para as camadas ocultas *relu* e *tanh*. Os melhores resultados foram alcançados com a função *tanh*;

c) *Funções de Otimização*: foram testados diferentes funções para retro-propagação dos gradientes mantendo os parâmetros de cada algoritmo, tais como a taxa de

Tabela 5.2: Parametrização para a Arquitetura Base LSTM.

Especialistas: Arquitetura Base LSTM		
Camada/Função	**Parâmetro**	**Valor**
Embedding Pré-treinado	Trainable	Static
	Fonte de Dados	Propósito Geral (6B)
	Input dimension	5000
	Output dimension	300
3 células LSTM	Units	16
	Activation Function	Tanh
	Dropout	0,20
	Return Sequence	True*
	Recurrent Dropout	0,20
	Kernel Initializer Function	Glorot Uniform
Saída	Units	3**
	Activation Function	Sigmoid
Modelo	Optimizer Function	Adam (learning = 0,001)
	Loss Function	Binary Crossentropy
Treinamento	Train Epochs	32
	Batch Size	40

* *False* para a última camada LSTM.

** Força a saída com três unidades para uso no modelo *DAC Stacking*.

aprendizado, conforme padrão fornecido pela biblioteca[7] usada. Os melhores resultados foram alcançados com a função *Adam*;

d) *Hiperparâmetros Específicos*: os experimentos exploraram os hiperparâmetros específicos *Return Sequence*, *Recurrent Dropout* e *Stateful*. A ativação do parâmetro *Return Sequence* em conjunto com o recurso *Recurrent Dropout* foi a que apresentou maior ganho de performance, sendo mantido para formação da arquitetura base;

e) *Hiperparâmetros Gerais e Parâmetros de Treinamento*: a partir da configuração base definida com os experimentos anteriores, variou-se cada parâmetro mantendo os demais fixos. Como configuração base dessa etapa, definiu-se o uso do *embedding* pré-treinado Glove 6B com aprendizado estático, e variou-se o total de camadas ocultas LSTM (2 a 4), unidades por camada (16 a 128), taxa de *dropout* (0,2 a 0,5), épocas de treinamento (16 a 100) e tamanho do lote (20 a 160). Ao final desse conjunto de experimentos, obteve-se a definição inicial para arquitetura LSTM.

Um último conjunto de experimentos envolveu o ajuste fino dos modelos LSTM. Nessa etapa, novamente foram exploradas variações para o número de neurônios por camada oculta, total de camadas LSTM consideradas, bem como os parâmetros de treinamento número de épocas e tamanho do lote.

5.5.1.2 Arquitetura CNN

Os experimentos realizados para a formação de modelos especialistas baseados na arquitetura CNN também exploram diferentes *embeddings* pré-treinados, funções de inicialização do *kernel* e hiperparâmetros específicos dessa arquitetura. Os principais aspectos desse conjunto de experimentos são destacados abaixo:

a) *Arquitetura base CNN*: o conhecimento obtido na etapa de treinamento dos modelos de arquitetura LSTM foi usado para fixar a configuração inicial dessa arquitetura. Assim, a arquitetura base CNN contempla o uso do *embedding* pré-treinado Glove 6B com aprendizado estático e função de inicialização do *kernel* Glorot Uniform. Em termos de camadas específicas, foram definidas uma camada convolucional 1D e uma camada *average pooling*, as quais são seguidas por uma camada de *dropout*

[7] https://keras.io/api/optimizers/

correspondente. Os hiperparâmetros gerais foram fixados para usar a função de ativação *relu* e a função de otimização *Ada Delta* (taxa de aprendizado igual à 0.001);

b) *Camada de embeddings e função de inicialização do kernel*: foram explorados tanto o uso de diferentes *embeddings* pré-treinados como o uso das funções Glorot e Lecun. O intuito desses experimentos é analisar a diferença de performance produzida pela mudança de arquitetura, considerando a variação dos mesmos recursos;

c) *Hiperparâmetros Específicos, Gerais e Parâmetros de Treinamento*. Como ajuste fino, ainda foram explorados os hiperparâmetros: total de filtros (100 e 250); tamanho do *kernel* (3 a 5) e taxa de *dropout* (20% e 50%). Para os parâmetros de treinamento foram variados o número de épocas de treinamento (10 e 20) e tamanho do lote (10, 20, 40). Essas variações foram aplicadas considerando diferentes *embeddings*.

A Tabela 5.3 apresenta o conjunto de parâmetros definidos para a topologia base CNN. A partir dessa topologia, os melhores resultados em termos de variação de hiperparâmetros específicos foram obtidos usando filtro de valor 250 para os classificadores de ansiedade e depressão, e 100 para o de comorbidade; tamanho do *kernel* igual a 5 para ansiedade e 4 para os classificadores de depressão e comorbidade; por fim, a taxa de *dropout* de 50% para os classificadores de ansiedade e comorbidade, e 20% para os especialistas em depressão.

Na segunda etapa de experimentos, foram explorados o uso de diferentes *embeddings* pré-treinados considerando a configuração base apresentada na Tabela 5.3. Para cada *embedding* pré-treinado e tipo de aprendizado (estático e não-estático), foram variados os valores dos hiperparâmetros *filters*, *kernel size* e taxa de *dropout*. Essas variações foram aplicadas para ajuste fino da arquitetura em relação aos *embeddings* testados.

5.5.1.3 Arquitetura Híbrida

A arquitetura base definida para o modelo híbrido é composta por uma rede CNN conectada a outra rede LSTM. A definição das camadas LSTM e CNN é baseada no conhecimento obtido durante o processo de formação dos modelos fracos detalhados nas seções 5.5.1.1 e 5.5.1.2, respectivamente. Assim, o número de camadas e o tipo de função de ativação em cada arquitetura específica foram mantidos, bem como o uso da função de inicialização do *kernel* (Glorot Uniforme). Para a função de otimização foi selecionado o

Tabela 5.3: Parametrização para a Arquitetura Base CNN

Camada/Função	Parâmetro	Valores		
		Ansiedade	**Depressão**	**Comorbidade**
Embedding Pré-treinado	Trainable	Static		
	Fonte de Dados	Propósito Geral Glove (6B)		
	Input dimension	5000		
	Output dimension	300		
1 célula CNN	Padding	Valid		
	Activation Function	ReLu		
	Kernel Initializer Function	Glorot Uniform		
	Filters	250	100	250
	Kernel Size	5	4	4
Global Average Pooling 1D	Data Format	Channel List		
Dropouts	Rate	0,5	0,2	0,5
Densa	Units	250	100	250
	Activation Function	ReLu		
Saída	Units	3*		
	Activation Function	Sigmoid		
Modelo	Optimizer Function	AdaDelta (learning = 0,001)		
	Loss Function	Binary Crossentropy		
Treinamento	Train Epochs	10		
	Batch Size	20		

* Força a saída com três unidades para uso no modelo *DAC Stacking*.

algoritmo *Ada Delta* (taxa de aprendizagem = 0,001). Para a camada de *embeddings* foi selecionado o Glove 6B parametrizado usando aprendizagem estática.

Os parâmetros explorados nos experimentos envolvendo os modelos híbridos são apresentados na Tabela 5.4. Desse conjunto de experimentos, destacam-se as variações:

a) *Hiperparâmetros da camada CNN*: a função de *pooling*, definida como *Max Pooling 1D*. Ainda, foram exploradas as variações para o número de filtros (32, 62 e 128); o tamanho do *kernel* (4, 5, 6) e a taxa de *dropout* (20% e 50%);

b) *Hiperparâmetros da camada LSTM*: foram explorados o total de unidades por camada (64, 128 e 256) e variações para a taxa de *dropout* (20% e 50%);

c) *Parâmetros de Treinamento*: Os treinamentos foram realizados fixando o tamanho do lote em 20 e variando o total de épocas de treinamento (10, 15, 20, 25 e 50).

5.5.2 Classificadores Diferenciadores entre Condições Mentais

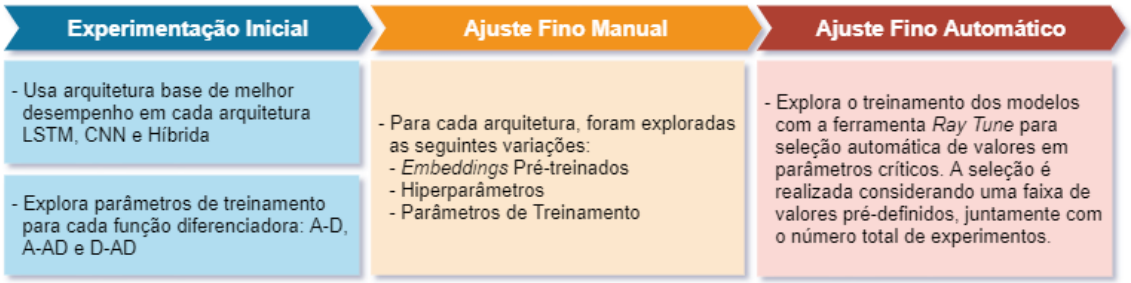
Os experimentos envolvendo a formação dos classificadores fracos diferenciadores é sintetizado na Figura 5.9. Esse fluxo de experimentos foi realizado para cada arquitetura, considerando os conjuntos de dados SMHD A-D, A-AD e D-AD (Tabela 4.3).

Tabela 5.4: Parametrização para a Arquitetura Base Híbrida

Camada/Função	Parâmetro	Valores		
		Ansiedade	**Depressão**	**Comorbidade**
Embedding Pré-treinado	Trainable	Static		
	Fonte de Dados	Propósito Geral Glove (6B)		
	Input dimension	5000		
	Output dimension	300		
1 célula CNN	Padding	valid		
	Activation Function	ReLu		
	Kernel Initializer Function	Glorot Uniform		
	Filters	128	128	64
	Kernel Size	5		
Max Pooling 1D	Data Format	channel list		
	Pool Size	2		
Dropouts CNN	Rate	0,2	0,5	0,5
3 células LSTM	Units	64		
	Activation Function	Tanh		
	Dropout	0,20		
	Return Sequence	True*		
	Recurrent Dropout	0,20		
	Kernel Initializer Function	Glorot Uniform		
Saída	Units	3*		
	Activation Function	Sigmoid		
Modelo	Optimizer Function	AdaDelta (learning = 0,001)		
	Loss Function	Binary Crossentropy		
Treinamento	Train Epochs	25	50	15
	Batch Size	20		

* Força a saída com três unidades para uso no modelo DAC Stacking.

Figura 5.9: Principais Etapas para formação do classificadores diferenciadores.



Os experimentos iniciais usaram como arquitetura base os modelos que apresentaram melhor performance na etapa de formação dos classificadores fracos especialistas, conforme detalhado na Seção 5.5.1. Esses experimentos exploraram o ajuste fino manual para os hiperparâmetros gerais e parâmetros de treinamento em cada arquitetura. De modo geral, os resultados obtidos permitiram constatar que desenvolver os modelos de classificação para diferenciar entre as condições mentais é uma tarefa computacionalmente muito mais difícil que identificar cada transtorno em relação aos respectivos usuários de controle. De modo geral, observou-se uma redução no desempenho desses modelos em relação aos classificadores especialistas. Considerando a medida F média,

por exemplo, verifica-se que a performance dos classificadores diferenciadores é inferior aos modelos especialistas em 19 pp.

Por essa razão, uma segunda etapa de experimentos foi realizada para seleção ótima de parâmetros com o uso do *framework Ray Tune*. Nessa etapa, os parâmetros e hiperparâmetros para cada arquitetura e função de classificação foram explorados segundo uma faixa de valores. As tabelas 5.5, 5.6 e 5.7 apresentam o conjunto de parâmetros e valores explorados para as arquiteturas LSTM, CNN e Híbrida, respectivamente.

Tabela 5.5: Diferenciadores: Arquitetura Base LSTM

***Ray Tune*: Parametrizações**		
Camada/Função	**Parâmetro**	**Valor(es)**
Embedding Pré-treinado	Fonte de Dados (Algoritmo, Aprenidizado)	Propósito Geral 6B (Glove, Estático)
		Targed Diagnosed Users (Glove, Estático)
		Targed Diagnosed Users (Word2Vec CBOW, Não-estático)
LSTM	Units	16 a 128
	Dropout	0,2 a 0,4
	Recurrent Dropout	0,2 a 0,4
	Kernel Initializer Function	LeCun Uniform
	Layers	3 a 5
Modelo	Learning Rate (Adam Optimizer Function)	0,001 a 0,01
Treinamento	Train Epochs	32 a 80
	Batch Size	10, 20, 25, 40

Tabela 5.6: Diferenciadores: Arquitetura Base CNN

***Ray Tune*: Parametrizações**		
Camada/Função	**Parâmetro**	**Valor(es)**
Embedding Pré-treinado	Fonte de Dados (Algoritmo, Aprenidizado)	Propósito Geral 6B (Glove, Estático)
		Propósito Geral Twitter (Glove, Estático)
		All Diagnosed Users (Word2Vec Skip-gram, Não-estático)
LSTM	Filters	200 a 512
	Dropout	0,2 a 0,6
	Kernel Size	3 a 6
	Kernel Initializer Function	Glorot Uniform
Densa	Units	200 a 512
Modelo	Learning Rate (AdaDelta Optimizer Function)	0,001 a 0,01
Treinamento	Train Epochs	10 a 50
	Batch Size	5, 10,20, 25, 40, 50

Tabela 5.7: Diferenciadores: Arquitetura Base Híbrida

***Ray Tune*: Parametrizações**		
Camada/Função	**Parâmetro**	**Valor(es)**
Embedding Pré-treinado	Fonte de Dados (Algoritmo, Aprenidizado)	Propósito Geral 6B (Glove, Estático) Propósito Geral Twitter (Glove, Estático) All Diagnosed Users (Word2Vec Skip-gram, Não-estático)
CNN	Filters	128 a 256
	Dropout	0,2 a 0,6
	Kernel Size	4 a 6
	Kernel Initializer Function	Glorot Uniform
LSTM	Units	64 a 256
	Dropout	0,2 a 0,35
	Recurrent Dropout	0,2 a 0,35
	Kernel Initializer Function	Glorot Uniform
Modelo	Learning Rate (AdaDelta Optimizer Function)	0,001 a 0,01
Treinamento	Train Epochs	25 a 55
	Batch Size	5, 10,20, 25, 40, 50

5.6 *Nível 1*: *Meta-learner*

O nível *meta-learner* é responsável por consolidar as predições do nível inferior em termos de um ou mais rótulos. Assim, essa camada necessita interpretar o contexto das predições fornecidas por cada classificador fraco e sua importância para a classificação final do usuário. As principais decisões envolvidas para formação desse nível compreendem a definição da (1) topologia do *Nível 0* em termos de composição para os classificadores fracos; (2) arquitetura para compor o modelo do nível *meta-learner*; e (3) abordagem empregada para o treinamento do modelo que compõe esse nível. O restante desta seção detalha essas escolhas.

5.6.1 Topologia dos Classificadores Fracos

As topologias para este nível variam de acordo com as propostas do *DAC Stacking* base (Figura 5.2) e suas variações *DAC Stacking EC* (Figura 5.3) e *DAC Stacking DT* (Figura 5.4). Em comum, todos apresentam em sua composição três classificadores fracos especialistas em ansiedade (CA_1, CA_2, CA_3), e três em depressão (CD_1, CD_2, CD_3). Na arquitetura *DAC Stacking EC* completam este nível três classificadores especialistas em comorbidade (CAD_1, CAD_2, CAD_3). Já no modelo *DAC Stacking DT*, são inclusos três classificadores diferenciadores (CA-D, CA-AD, CD-AD).

Cada conjunto de classificadores fracos de mesmo objetivo apresenta variações quanto à arquitetura e/ou *embedding* empregados. A escolha de três classificadores para cada grupo visa forçar uma decisão de classificação, considerando cada condição alvo.

5.6.2 Arquitetura do Modelo *Meta-learner*

Para compor o modelo do nível *meta-learner*, optou-se por uma rede neural densa, composta de diferentes camadas *perceptron* totalmente conectadas. O número de camadas ocultas, unidades por camada, tamanho do lote e total de épocas de treinamento, bem como os hiper-parâmetros dessa rede neural foram definidos experimentalmente usando como topologia o modelo *DAC Stacking* (base) apresentado em (SOUZA; NOBRE; BECKER, 2020).

O modelo *meta-learner* de melhor desempenho foi aquele composto por três camadas Densas. Cada camada Densa foi definida para usar a função de ativação *tanh* e 12 unidades dedicadas a processar as predições de cada classificador fraco presente no *Nível 0*. Para a camada de saída, foram mantidos os mesmos parâmetros definidos na camada de saída dos classificadores fracos. A Tabela 5.10 apresenta a parametrização final para a rede neural densa do *Nível 1*, a qual foi mantida em todas as variações propostas para o comitê *DAC Stacking*.

Figura 5.10: *Meta-learner*: Configuração da Arquitetura Base Rede Neural Densa

Camada/Função	Parâmetro	Valor(es)
Densa	Hidden Layers	3
	Units by model/Layer*	12
	Activation Function	tanh
	Kernel Initializer Function	Glorot Uniform
Saída	Units	3
	Activation Function	Sigmoid
Modelo	Optimizer Funciton	Adam (learning rate = 0.001)
	Loss Function	binary crossentropy
Treinamento	Train Epochs	16
	Batch Size	8

* O número de unidades dedicadas para cada classificador fraco.

5.6.3 Estratégias de Treinamento

A definição da abordagem para o treinamento do modelo *meta-learner* envolveu uma questão importante acerca dos dados usados nesse processo. Durante a etapa de treinamento dos classificadores fracos, foram utilizadas as divisões de treinamento e validação dos respectivos conjuntos de dados de rótulo único mostrados na Tabela 4.3. Essa estratégia foi adotada sobre a premissa de que cada classificador fraco não deve ser influenciado pelos resultados gerados por outros classificadores individuais e, portanto, deve ser treinado independentemente um do outro.

De modo similar, a rede neural densa do *Nível 1* deve ser treinada usando um conjunto de instâncias até então desconhecidas para os classificadores fracos, a fim de não introduzir viés nos resultados. Por essa razão, o modelo do *Nível 1* foi treinado usando a divisão de teste do conjunto de original SMHD A-D-AD, redistribuído em uma proporção de 80% para treinamento/validação e 20% para teste do modelo final. Para compensar o menor número de instâncias de treinamento, em comparação com os conjuntos usados para treinar os classificadores fracos, adotou-se o método de validação cruzada. Visto que cada execução pode levar muitas horas, optou-se por um *k-fold* = 5.

5.7 Avaliação de Desempenho

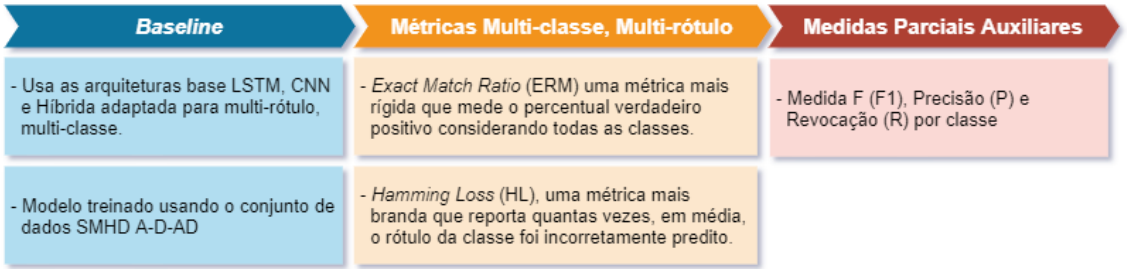
Para a avaliação de desempenho dos modelos *DAC Stacking* e suas variações propõe-se um método que abrange tanto a análise quantitativa, quanto qualitativa desses modelos. O restante desta seção é dedicado à descrição detalhada desses métodos.

5.7.1 Método para Avaliação Quantitativa

A Figura 5.11 sintetiza os principais aspectos do método para a avaliação quantitativa dos modelos *DAC Stacking*. O método envolveu as seguintes definições:

a) *Baseline*: como mencionado anteriormente, raros trabalhos abordam a classificação de usuários em condição de comorbidade. A única proposta de modelo multi-tarefa no mesmo conjunto de dados (COHAN et al., 2018) apresentou resultado insatisfatório, como destacado na seção 3.4. Além do mais, embora seja dedicada à classificação incluindo condições de comorbidade a partir da rede social Reddit, aqueles

Figura 5.11: Etapas do Método de Avaliação Quantitativa



modelos não são diretamente comparáveis, uma vez que os autores exploram outras condições mentais além dos transtornos alvo desse estudo.

Na ausência de *baselines*, os classificadores *DAC Stacking* foram comparados com os modelos multirrótulo de arquiteturas LSTM, CNN e Híbrida, treinados usando o conjunto de dados SMHD A-D-AD da Tabela 4.3. Para definição desses *baselines*, adaptou-se as arquiteturas base LSTM, CNN e Híbrida de melhor desempenho nos classificadores fracos especialistas;

b) *Métricas para modelos multi-tarefa*: para a avaliação quanto ao desempenho na tarefa de classificação multirrótulo, adotou-se as métricas *Exact Match Ratio* e *Hamming Loss*. Estas métricas permitem avaliar a performance quanto à multi-tarefa, mas elas não mostram diretamente a performance em relação a cada transtorno alvo;

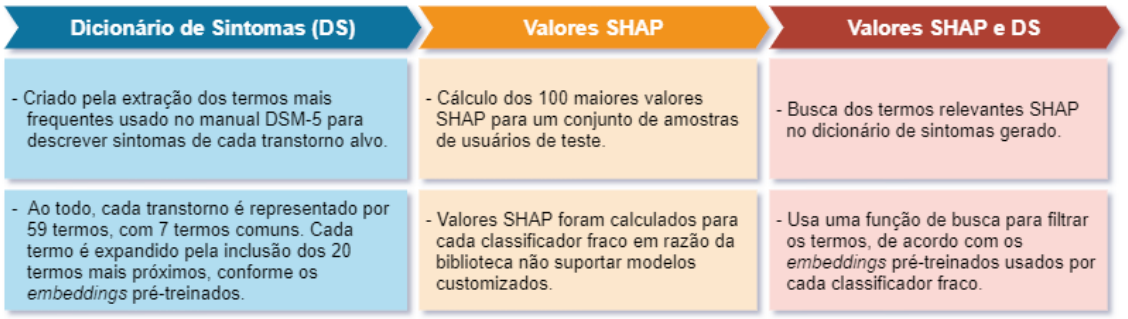
c) *Métricas para modelos binários*: envolve a definição de métricas auxiliares para análise individual de performance considerando cada condição alvo. Para tanto, adotou-se as métricas tradicionais empregadas para avaliação de modelos binários: medida F, Precisão e Revocação, as quais são calculadas considerando cada classe.

5.7.2 Método para Avaliação Qualitativa

A análise qualitativa visa compreender se as *features* mais influentes para a classificação, segundo cada classificador fraco presente no *Nível 0*, são representativas dos sintomas conhecidos para cada condição alvo. Para tanto, o método proposto para realização dessa análise compreende três etapas, sumarizadas na Figura 5.12.

Inicialmente, foram compostos Dicionários de Sintomas (DS) para cada condição alvo. O processo de formação dos mesmos é sintetizado na Figura 5.13. Primeiramente, foram extraídos os termos mais frequentes contidos na descrição dos transtornos de an-

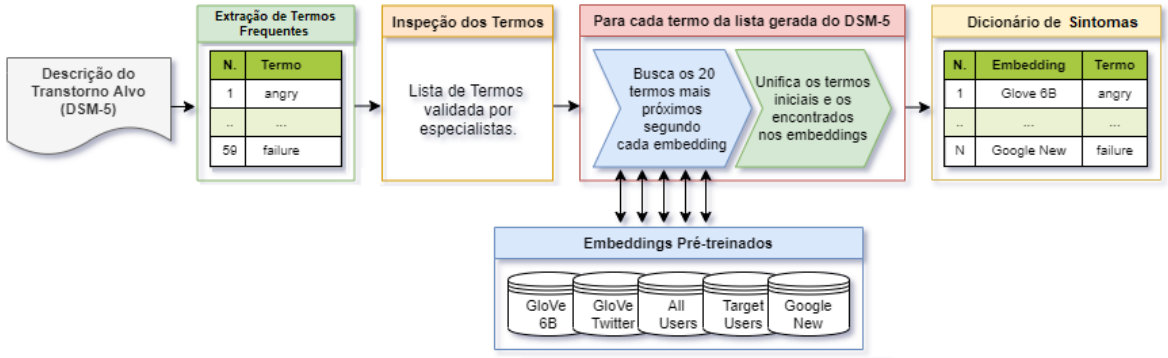
Figura 5.12: Fluxo para análise de *features* influentes na classificação.



siedade e depressão, segundo o manual DSM-5, resultando em uma lista de 938 termos. Essa lista de termos foi validada por dois especialistas com formação em psicologia, de acordo com o critério de seleção sintomas mais frequentes para cada condição alvo. Ao final dessa validação, cada transtorno foi representado por uma lista de 59 termos, sendo 7 comuns entre eles. A Figura 5.14 apresenta uma nuvem de palavras com os termos mais frequentes para os transtornos de depressão e ansiedade. A relação completa de termos está disponível no Apêndice C. Os termos encontrados no DSM-5 representam os sintomas segundo uma linguagem técnica, a qual dificilmente é usada por usuários em redes sociais. Por esta razão, cada termo da lista inicial foi expandido pelo acréscimo dos 20 termos mais próximos a ele, segundo os diferentes *embeddings* usados para construir os classificadores fracos.

Na segunda etapa, foram calculados os 100 valores SHAP mais altos para um conjunto de amostras de teste, correta e incorretamente classificadas. Para este fim, adotamos a biblioteca *Kernel Explainer*[8]. Esses valores são calculados para cada classificador fraco do *Nível 0*, já que esta biblioteca não suporta o modelo do tipo comitê.

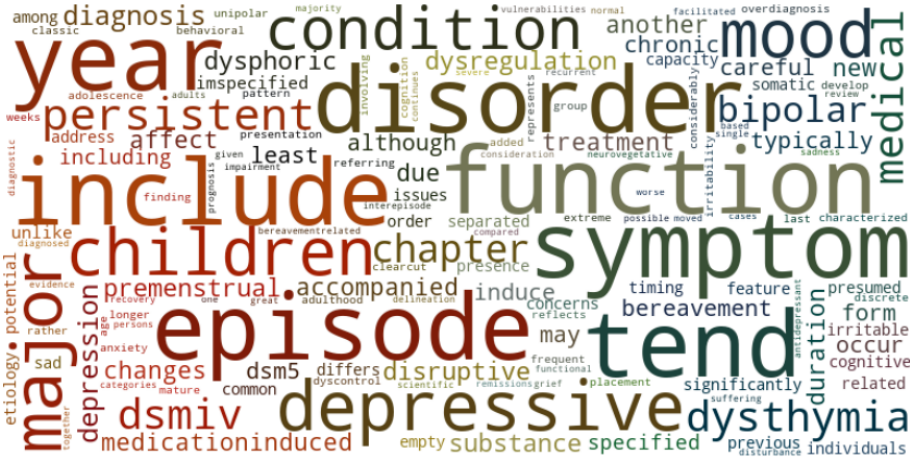
Figura 5.13: Processo de Formação dos Dicionários de Sintomas.



A etapa final do método proposto consiste em analisar a relação entre os termos

[8]https://shap.readthedocs.io/en/latest/#shap.KernelExplainer

Figura 5.14: DSM-5: Termos mais frequentes para a descrição dos transtornos alvo



(a) Transtornos Depressivos

(b) Transtornos de Ansiedade

relevantes para a decisão final de classificação do modelo *DAC Stacking*, segundo valores SHAP, e os possíveis sintomas para cada condição alvo. Para tanto, definiu-se uma função de busca que permite filtrar esses termos nos dicionários de sintomas, conforme o *embedding* pré-treinado usado por cada classificador fraco que compõe o *Nível 0* do *DAC Stacking*. Assim, é realizada uma análise considerando o conjunto de termos retornado por essas buscas e sua relação com a descrição de sintomas relatados no DSM-5, para obtenção de *insights* acerca dos padrões identificados para cada condição alvo.

6 EXPERIMENTOS E RESULTADOS

Este capítulo apresenta os experimentos realizados para o desenvolvimento e avaliação dos modelos *DAC Stacking* e suas variações, propostas no Capítulo 5.

6.1 Objetivos

Os experimentos propostos visaram responder as questões de pesquisa formuladas neste trabalho, detalhados no Capítulo 1. Para tanto, esses experimentos foram realizados segundo os objetivos detalhados abaixo:

- *Experimento #1*: verificar se alguma dentre as arquiteturas de redes neurais testadas (LSTM, CNN e Híbrida) apresenta melhor desempenho para classificação das condições alvo. Para tanto, esses experimentos foram organizados para explorar a formação dos classificadores fracos a partir de cada tipo de arquitetura, considerando as funções de especialista e diferenciadores entre condições alvo;
- *Experimento #2*: verificar o impacto da variação de *embeddings* pré-treinados de propósito geral e de domínio no desempenho dos classificadores fracos, de acordo com arquiteturas das redes neurais testadas;
- *Experimento #3*: avaliar quantitativamente o desempenho do modelo *DAC Stacking* e suas variações (*DAC Stacking EC* e *DAC Stacking DT*) na classificação de ansiedade, depressão e condição de comorbidade segundo um problema multirrótulo;
- *Experimento #4*: avaliar qualitativamente o desempenho para modelo *DAC Stacking* de melhor performance, quanto à relação entre as *features* influentes para a classificação das instâncias por esse modelo e os sintomas apresentados para cada condição alvo;
- *Experimento #5*: experimento adicional para avaliar a performance do modelo *DAC Stacking* proposto em relação à arquitetura BERT, atual estado da arte para problemas de classificação em PLN.

Os experimentos propostos foram executados usando os conjuntos de dados derivados do SMHD (Tabela 4.3), conforme procedimentos descritos no Capítulo 5. Vale lembrar que, as instâncias destes conjuntos de dados são divididas em proporções iguais

para formar os subconjuntos de treinamento, validação e teste. Todos os experimentos incluíram o recurso de monitoramento de performance do modelo, o qual analisa o desempenho deste ao final de cada época de treinamento, permitindo salvar a versão de melhor desempenho do modelo treinado. Esse monitoramento considerou o máximo valor alcançado para a métrica acurácia. O restante deste capítulo detalha cada experimento e seus resultados.

6.2 Experimento #1: Formação dos Classificadores Fracos

Como detalhado na Seção 5.5, esses modelos foram definidos de forma empírica, como resultado de um extenso processo de experimentação envolvendo variações de hiperparâmetros e *embeddings* pré-treinados. Neste experimento, analisou-se os resultados em termos de desempenho de acordo com cada arquitetura usada no desenvolvimento dos classificadores especialistas e diferenciadores, segundo as diferentes arquiteturas exploradas para cada condição alvo. Esses experimentos visam responder em parte à questão de pesquisa QP2, i.e "*As distintas premissas de aprendizado, subjacentes às arquiteturas de aprendizado profundo exploradas (padrões locais, sequenciais), contribuem de forma isolada ou combinada à melhoria de desempenho da classificação em uma abordagem stacking ensemble?*". Os melhores modelos resultantes desse conjunto de experimentos foram usados para compor o *Nível 0* dos modelos *DAC Stacking* e suas variações.

6.2.1 Método

Para avaliar a performance das arquiteturas LSTM, CNN e Híbrida para a formação de classificadores fracos, o experimento foi organizado como descrito abaixo:

- *Especialistas*: os classificadores binários dedicados à identificação da ansiedade, depressão e comorbidade foram desenvolvidos usando as instâncias de treinamento e validação dos conjuntos de dados SMHD A, SMHD D e SMHD AD, respectivamente (Tabela 4.3). Para a formação de cada classificador especialista, foram exploradas as arquiteturas LSTM, CNN e Híbrida, variando os hiperparâmetros específicos de cada arquitetura e parâmetros de treinamento, como detalhado na Seção 5.5.1;

- *Diferenciadores*: os classificadores binários projetados para diferenciar entre as

condições alvo, foram desenvolvidos usando as instâncias de treinamento e validação dos conjuntos de dados SMHD A-D, SMHD A-AD e SMHD D-AD (Tabela 4.3). Conforme detalhado na Seção 5.5.2, as parametrizações definidas para as arquiteturas LSTM e CNN durante a formação dos classificadores especialistas foram usadas como base para esse conjunto de experimentos. Os resultados iniciais mostraram que a diferenciação entre as condições alvo apresenta padrões mais sutis, o que motivou o método alternativo de experimentação para variação dos hiperparâmetros em cada arquitetura, já apresentado na Seção 5.5.2. Este explorou tanto o ajuste fino manual, quanto automático com o uso da ferramenta *Ray Tune*. Em relação às arquiteturas, somente foram exploradas as redes neurais LSTM e CNN. A arquitetura híbrida apresentou custo computacional proibitivo, associado a uma performance insatisfatória, tanto no ajuste manual, quanto no ajuste automático, sendo abandonado para formação dos modelos diferenciadores.

A análise de performance dos classificadores fracos foi realizada considerando a média dos resultados ao longo de 10 repetições, gerada a partir da topologia de melhor desempenho em cada condição alvo, segundo cada arquitetura. Com isso, tem-se como variável apenas a inicialização randômica dos pesos da rede neural a cada repetição do experimento. O desempenho de cada modelo foi mensurado usando as métricas de avaliação para os classificadores binários Precisão (P), Revocação (R) e medida F (F1) extraídas segundo a performance apresentada por esses modelos em relação ao respectivo conjunto de teste. Para cada arquitetura, foram selecionados os modelos de melhor desempenho, ordenados segundo a medida F. Para os modelos com performance semelhantes para a medida F, adotou-se como critérios de seleção complementar o melhor equilíbrio entre as métricas de precisão e revocação, e/ou o emprego de diferentes *embeddings*, visando a ganho de variabilidade para a composição do *Nível 0* do modelo *DAC Stacking*.

Para realizar a análise de significância estatística dos resultados foi adotado o Teste Student T bicaudal pareado ($alpha = 0.5$) (SPIEGEL et al., 2009). Essa análise consiste em, para cada condição alvo, comparar a performance das métricas de avaliação entre duas arquiteturas. Como hipótese nula, considerou-se que para uma mesma condição alvo não há diferença entre as duas arquiteturas comparadas. Como hipótese alternativa ($p - value < 0.05$), tem-se que existe diferença de performance entre as arquiteturas analisadas para a tarefa de classificar uma condição alvo. Por fim, a curva ROC foi usada para estabelecer um comparativo de performance entre as diferentes arquiteturas selecionadas para os classificadores especialistas, segundo cada condição alvo.

As seções seguintes apresentam os resultados separadamente, e uma discussão comparativa entre os experimentos realizados para formação dos classificadores especialistas e diferenciadores.

6.2.2 Classificadores Especialistas

Os experimentos realizados para o desenvolvimento dos classificadores especialistas foram analisados quanto à performance por arquitetura e condição alvo, com o intuito de selecionar os três modelos de melhor desempenho segundo esses critérios. As Tabelas 6.1, 6.2 e 6.3 apresentam os classificadores fracos com melhor desempenho para as arquiteturas LSTM, CNN e Híbrida, respectivamente, os quais foram selecionados para a composição do *Nível 0* de modelos *DAC Stacking*. Para cada condição alvo, destaca-se em negrito o melhor resultado em termos de medida F. Demais resultados usados para análise estatística dos modelos gerados são detalhados no Apêndice D.

Cada tabela apresenta os resultados para os modelos selecionados, destacando as variações para a topologia em relação a cada arquitetura base. Para distinguir esses modelos, adotou-se uma convenção de nomenclatura que permite identificar sua arquitetura e função. Assim, os classificadores fracos foram nomeados segundo um prefixo que representa a inicial de cada arquitetura que compõe o modelo, seguidos da identificação de sua função. Por exemplo, para os classificadores fracos especialistas em ansiedade, formados pela arquitetura LSTM, tem-se o nome *L-CA*$_i$, onde *i* representa o modelo selecionado conforme as demais variações (*embedding* e parametrizações de cada rede neural).

Por fim, a formação dos classificadores fracos destacados nessas tabelas também considerou experimentos relacionados à variação da função *kernel* de inicialização da rede, cujos resultados são apresentados no Apêndice B. Esses experimentos permitiram concluir que a variação dessa função impacta na performance somente dos modelos de arquitetura LSTM, estabelecendo um *trade-off* entre revocação e precisão.

6.2.2.1 Resultados

a) Arquitetura LSTM. Como mostra a Tabela 6.1, de modo geral, essa arquitetura apresentou uma performance para medida F entre 0,65 e 0,77. Para os modelos especialistas em uma condição alvo isolada (ansiedade ou depressão), o melhor desempenho foi observado na topologia que inclui o uso do *embedding* pré-treinado Glove 6B e apren-

dizado estático. Para os modelos especialistas em comorbidade, o melhor desempenho foi observado usando o *embedding* pré-treinado de domínio *All Diagnosed Users*, com aprendizado não estático.

Tabela 6.1: Performance Final do Classificadores Binários LSTM

Condições Mentais	Modelo	Embedding	Algoritmo	Aprendizado	Kernel Initializer	P	R	F1
Ansiedade	L-CA1	**General Purpose (6B)**	**Glove**	**Estático**	Glorot Uniform	0,73	0,62	0,67
	L-CA2				**LeCun Uniform**	**0,73**	**0,69**	**0,71**
	L-CA3	Targed Diagnosed Users*			Glorot Uniform	0,62	0,79	0,70
Depressão	L-CD1	**General Purpose (6B)**	**Glove**	**Estático**	Glorot Uniform	0,75	0,77	0,76
	L-CD2				**LeCun Uniform**	**0,74**	**0,79**	**0,77**
	L-CD3	Targed Diagnosed Users*	Word2Vec CBOW	Não-estático	Glorot Uniform	0,67	0,81	0,73
Comorbidade	L-CAD1	General Purpose (6B)	Glove	Estático	Glorot Uniform	0,72	0,58	0,65
	L-CAD2				LeCun Uniform	0,67	0,64	0,66
	L-CAD3	**All Diagnosed Users***	**Word2Vec CBOW**	**Não-estático**	**Glorot Uniform**	**0,77**	**0,67**	**0,72**

* Variações em relação à arquitetura LSTM base: Batch size = 20, Train Epochs = 96, Dropout = Recurrent Dropout = 0,1.

Analisando a performance para cada condição mental, verifica-se que os modelos dedicados à depressão apresentaram em média performance melhor em todas as métricas. Em termos de medida F, os classificadores de depressão foram superiores aos classificadores de ansiedade em 6 pp, e aos de comorbidade em 8 pp.

Um desempenho médio similar foi observado em termos de medida F ao comparar os classificadores de ansiedade e comorbidade. No entanto, os modelos apresentam comportamento inverso em relação às métricas de precisão e revocação, onde foi observado que os classificadores de ansiedade alcançaram ganhos maiores para a revocação, em detrimento da precisão.

b) Arquitetura CNN. Os modelos formados a partir da arquitetura CNN detalhados na Tabela 6.2, apresentaram a melhor performance entre as arquiteturas exploradas. Em termos de medida F1, o desempenho desses modelos foi similar em todas as condições alvo, variando entre 0,76 e 0,79. Em relação às demais arquiteturas, as maiores diferenças foram notadas para os modelos dedicados à comorbidade, onde a performance média para a medida F1 foi superior 11 pp em relação aos classificadores baseados na arquitetura LSTM, e 14 pp em relação ao modelos híbridos.

De modo geral, os modelos CNN apresentaram performance similares para os diferentes *embeddings* pré-treinados. Assim, a seleção dos melhores resultados da arquitetura CNN considerou os modelos cujos *embeddings* não estão presentes nos modelos de

Tabela 6.2: Arquitetura CNN: Performance Final dos Classificadores Especialistas

Condições Mentais	Modelo*	Embedding	Algoritmo	Variações de Configuração			P	R	F1
				Filters	Kernel Size	Dropout			
Ansiedade	C-CA1 (base)	General Purpose (6B)	Glove	250	5	0,5	0,77	0,75	0,76
	C-CA2	General Purpose (Twitter)	Glove	100	5	0,5	0,78	0,77	0,78
	C-CA3	**All Diagnosed Users**	**Word2Vec Skip-gram**	**250**	**4**	**0,2**	**0,72**	**0,85**	**0,78**
Depressão	C-CD1 (base)	General Purpose (6B)	Glove	250	4	0,2	0,76	0,81	0,79
	C-CD2	**General Purpose (Twitter)**	**Glove**	**250**	**3**	**0,5**	**0,72**	**0,87**	**0,79**
	C-CD3	Targed Diagnosed Users	Word2Vec Skip-gram	100	3	0,2	0,76	0,82	0,79
Comorbidade	**C-CAD1 (base)**	**General Purpose (Google News)**	**Word2Vec**	**100**	**5**	**0,5**	**0,74**	**0,84**	**0,79**
	C-CAD2	General Purpose (Twitter)	Glove	250	3	0,5	0,74	0,84	0,78
	C-CAD3	Targed Diagnosed Users	Word2Vec Skip-gram	250	5	0,2	0,77	0,79	0,78

* Todos os modelos apresentaram melhor performance com aprendizado estático.

Tabela 6.3: Arquitetura Híbrida: Performance Final dos Classificadores Especialistas

Condições Mentais	Modelo*	Células CNN			Células LSTM		P	R	F
		Filters	Kernel Size	Dropout	Unidades	Dropout			
Ansiedade	**H-CA1 (base)**	**128**	**5**	**0,2**	**64**	**0,2**	**0,71**	**0,78**	**0,74**
	H-CA2	128	4	0,2	64	0,2	0,77	0,71	0,74
	H-CA3	32	5	0,2	256	0,2	0,67	0,71	0,69
Depressão	H-CD1 (base)	128	5	0,5	128	0,2	0,73	0,80	0,76
	H-CD2	**128**	**5**	**0,5**	**128**	**0,2**	**0,81**	**0,73**	**0,77**
	H-CD3	128	5	0,5	128	0,2	0,70	0,73	0,71
Comorbidade	H-CAD1 (base)	32	5	0,2	256	0,2	0,59	0,58	0,58
	H-CAD2	**64**	**5**	**0,5**	**256**	**0,5**	**0,70**	**0,74**	**0,72**
	H-CAD3	64	5	0,2	256	0,2	0,53	0,80	0,64

* Todos os modelos usam \ *embedding* de propósito geral Glove 6B com aprendizado estático.

arquitetura LSTM, com a finalidade de contribuir para a variabilidade entre os classificadores fracos usados para a composição do *DAC Stacking*.

c) Arquitetura Híbrida. Como mostra a Tabela 6.3, os melhores modelos híbridos apresentaram performance média para a medida F1 entre 0,64 e 0,75. Similar ao comportamento observado nos classificadores baseados na arquitetura LSTM, os modelos híbridos também apresentaram melhor performance para a função especialista em depressão. Em termos de performance média F1, os classificadores de depressão foram 2 pp superiores aos de ansiedade, e 10 pp aos de comorbidade. No entanto, diferente do observado para a arquitetura LSTM, os classificadores de ansiedade apresentaram maior equilíbrio entre as métricas precisão e revocação (diferença de 1 pp). Já os classificadores de comorbidade apresentaram uma diferença mais acentuada entre essa métricas (10 pp).

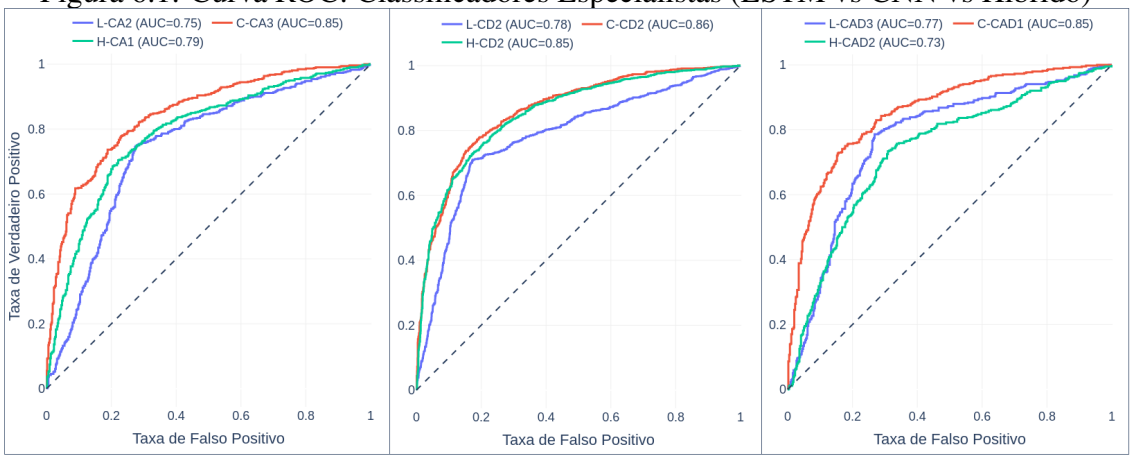
Todos os modelos apresentados na Tabela 6.3 foram gerados usando o *embedding* pré-treinado Glove 6B considerando o aprendizado estático. Cabe destacar que os mode-

los híbridos foram explorados segundo um conjunto menor de experimentos, devido ao conhecimento prévio sobre o comportamento de cada arquitetura, adquirido nas etapas de testes anteriores. Desse modo, os *embeddings* pré-treinados não foram testados de modo exaustivo. Assim, o conjunto de experimentos explorou a arquitetura base quanto aos *embeddings* pré-treinados (1) Glove 6B e Glove Twitter com aprendizado estático; e (2) *All Diagnosed User* (CBOW) com aprendizado não estático.

6.2.2.2 Discussão

A Figura 6.1 apresenta as curvas ROC geradas para os modelos de melhor performance em cada arquitetura, considerando as diferentes condições alvo. De modo geral, observa-se que os classificadores baseados na arquitetura CNN apresentaram a melhor performance para todas as condições alvo. Em termos de medida AUC, esse desempenho é muito similar entre os especialistas, sendo 0,86 para depressão e 0,85 para os classificadores de ansiedade e comorbidade.

Figura 6.1: Curva ROC: Classificadores Especialistas (LSTM vs CNN vs Híbrido)



(a) Ansiedade (b) Depressão (c) Comorbidade

Em relação aos classificadores baseados nas arquiteturas LSTM e Híbrida, observa-se um comportamento que difere conforme a condição alvo. Considerando a medida AUC, o modelo híbrido foi superior ao LSTM em 4 pp para a função de especialista em ansiedade, como mostra a Figura 6.1(a). Essa mesma diferença é observada na Figura 6.1(c) em favor do modelo LSTM para os especialistas em comorbidade.

Analisando a performance por condição alvo, nota-se que os modelos dedicados à identificação da depressão apresentam os melhores ganhos de performance. Comparando a curva ROC entre esses classificadores especialistas, observa-se um desempenho muito

similar entre os modelos baseados nas arquiteturas CNN e Híbrida, como apresenta a Figura 6.1(b).

Por fim, os modelos gerados para as três arquiteturas são comparáveis ao estado da arte para a classificação das condições alvo. Considerando a medida F para os especialistas em depressão, os resultados médios dos modelos selecionados foram 0,79 para arquitetura CNN e 0,75 para as arquiteturas LSTM e Híbrida. Esses resultados são superiores ao modelo (YATES; COHAN; GOHARIAN, 2017) e comparáveis aos modelos (TADESSE et al., 2019a; MANN; PAES; MATSUSHIMA, 2020). Em relação aos classificadores de ansiedade, os modelos de arquitetura CNN apresentam performance comparável, em termos de precisão e revocação, com os modelos apresentados em Shen e Rudzicz (2017). Para a comorbidade, não foram encontrados trabalhos focados na identificação dessa condição usando modelos binários.

6.2.3 Classificadores Diferenciadores

Os experimentos realizados para a formação dos classificadores diferenciadores exploraram apenas as arquiteturas LSTM e CNN para a formação dos modelos A-D, A-AD, e D-AD. Os melhores resultados desse conjunto de experimentos são apresentados na Tabela 6.4. De modo similar à nomenclatura definida para os classificadores especialistas, os modelos diferenciadores também foram nomeados visando a identificação da arquitetura e condições alvo em que foram treinados. Assim, para os classificadores diferenciadores entre ansiedade e depressão compostos pela arquitetura LSTM, por exemplo, tem-se o nome *L-CA-*D_i, onde i representa as diferenças entre os modelos selecionados, conforme demais parametrizações da rede neural.

No restante desta seção, um comparativo entre as abordagens de treinamento por ajuste manual e ajuste automático é apresentado nas Figuras 6.2, 6.3 e 6.4 para os classificadores A-D, A-AD e D-AD, respectivamente. Esse comparativo considera a distribuição de performance da medida F para todo o conjunto de modelos deste experimento.

6.2.3.1 Resultados

Os melhores resultados para cada função diferenciadora estão especificados na Tabela 6.4. De modo geral, os classificadores diferenciadores apresentaram uma performance bem inferior quando comparada a dos modelos especialistas, discutidos na se-

Tabela 6.4: Classificadores Diferenciadores: Modelos de melhor Performance.

Codições Mentais	Modelo Diferenciador	*Embedding*	Algoritmo	Aprendizado	Primeira Condição Mental			Segunda Condição Mental		
					P	R	F1	P	R	F1
Ansiedade e Depressão	L-CA-D1	Propósito Geral 6B	Glove	Estático	0,56	0,53	0,55	0,56	0,57	0,56
	C-CA-D2	**Target Diagnosed Users**	**Word2Vec CBOW**	**Não-Estático**	**0,56**	**0,56**	**0,56**	**0,56**	**0,55**	**0,56**
	C-CA-D3	Propósito Geral Twitter	Glove	Estático	0,56	0,62	0,59	0,57	0,51	0,54
Ansiedade e Comorbidade	**L-CA-AD1**	**Propósito Geral 6B**	**Glove**	**Estático**	**0,57**	**0,56**	**0,56**	**0,57**	**0,58**	**0,57**
	C-CA-AD2	Target Diagnosed Users	Word2Vec CBOW	Não-Estático	0,57	0,65	0,60	0,59	0,50	0,54
	C-CA-AD3	Propósito Geral Twitter	Glove	Estático	0,56	0,65	0,61	0,59	0,50	0,54
Depressão e Comorbidade	L-CD-AD1	Propósito Geral 6B	Glove	Estático	0,52	0,52	0,52	0,52	0,53	0,53
	C-CD-AD2	Target Diagnosed Users	Word2Vec CBOW	Não-Estático	0,52	0,55	0,53	0,52	0,50	0,51
	C-CD-AD3	**Propósito Geral Twitter**	**Glove**	**Estático**	**0,53**	**0,55**	**0,54**	**0,54**	**0,52**	**0,53**

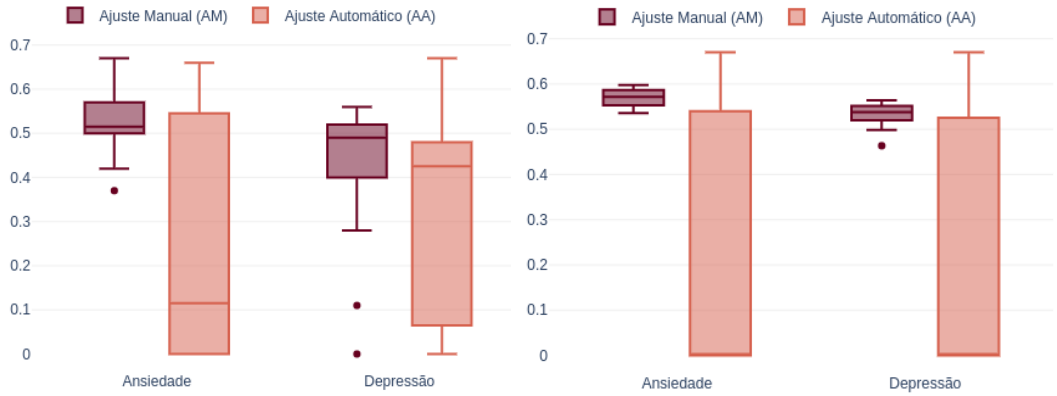
ção anterior. Considerando a performance média da medida F por tipo de diferenciador, observam-se resultados distintos. Os classificadores diferenciadores D-AD apresentaram as menores médias para medida F, sendo 53% para a depressão e 52% para a comorbidade. Para os modelos diferenciadores A-AD, a performance média atingida para cada condição foi maior, sendo 59% para ansiedade e 55% para a comorbidade. Por fim, os modelos diferenciadores A-D apresentaram o desempenho médio inverso para as condições de ansiedade e depressão, quando comparados aos classificadores A-AD e D-AD, respectivamente. Uma redução de 2 pp foi observado para a identificação da ansiedade em comparação ao modelo A-AD, e na mesma proporção, um aumento para identificação da depressão é observado em relação aos diferenciadores D-AD.

Os resultados também foram avaliados em termos da distribuição de performance para a medida F, estabelecendo um comparativo entre as arquiteturas LSTM e CNN, considerando tanto o ajuste fino manual (AM), quanto o automático (AA). Os resultados desse comparativo são explanados abaixo, para cada tipo de modelo diferenciador.

a) Diferenciadores de Ansiedade e Depressão (A-D). A Figura 6.2 apresenta o comparativo para a função de diferenciação entre Ansiedade e Depressão (A-D). Verifica-se que os modelos gerados por ajuste manual apresentaram em média performance superior. Como esperado, o ajuste automático apresentou uma variação maior de performance. Para os modelos CNN, mostrados na Figura 6.2 (a), os melhores resultados foram produzidos por ajuste automático. Já os modelos LSTM apresentaram maior equilíbrio de performance entre as classes com ajuste manual, como mostra Figura 6.2 (b).

b) Diferenciadores de Ansiedade e Comorbidade (A-AD). O comparativo de per-

Figura 6.2: Diferenciadores A-D: Performance F1 (AM vs AA)

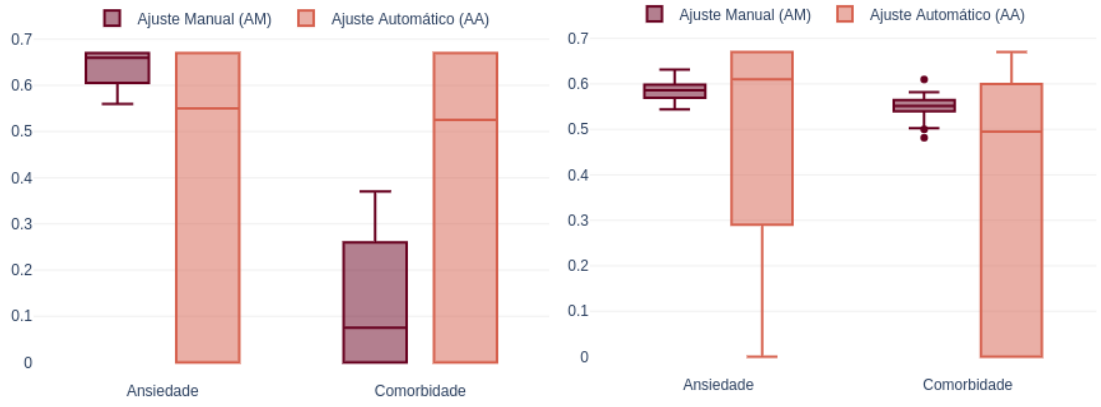


(a) Modelos LSTM (b) Modelos CNN

formance entre as arquiteturas e abordagens de treinamento, considerando os classificadores diferenciadores entre as condições Ansiedade e Comorbidade é apresentado na Figura 6.3. Observa-se na Figura 6.3 (a) que, os modelos LSTM desempenharam melhor em média para identificação da ansiedade, em detrimento da comorbidade, considerando o ajuste manual. Para o ajuste automático, verifica-se que a performance para a medida F em cada classe foi mais equilibrada, porém inferior à média atingida com ajuste manual. Um comportamento inverso é observado ao comparar a performance por classe e tipo de ajuste, considerando a arquitetura CNN, como mostra a Figura 6.3 (b).

c) Diferenciadores de Depressão e Comorbidade (D-AD). Por fim, o comparativo considerando os classificadores diferenciadores entre as condições Depressão e Comorbidade é apresentado na Figura 6.4. Ao comparar os experimentos LSTM com CNN, considerando o ajuste automático, verifica-se um comportamento médio inverso para a performance em cada classe. Nos modelos LSTM, a comorbidade foi beneficiada, enquanto que nos modelos CNN a Depressão desempenhou melhor. Considerando o ajuste

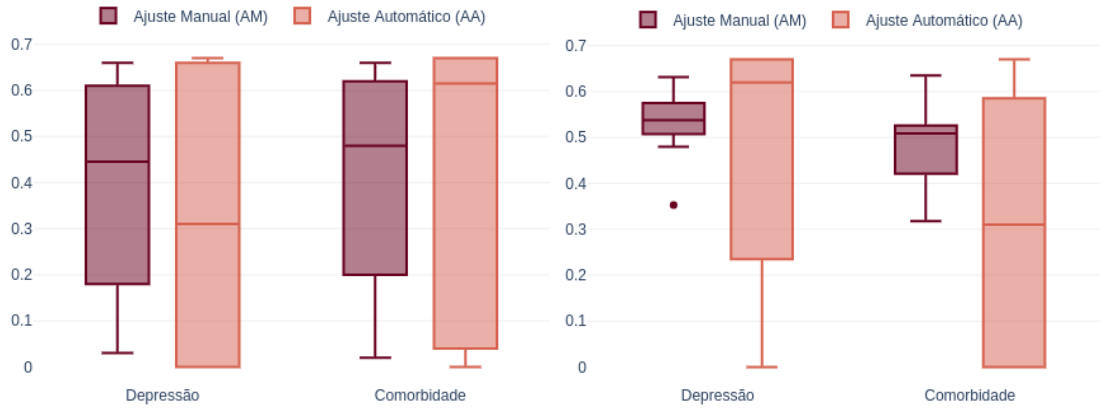
Figura 6.3: Diferenciadores A-AD: Performance F1 (AM vs AA)



(a) Modelos LSTM (b) Modelos CNN

manual para experimentos em ambas arquiteturas, observa-se um equilíbrio maior de performance entre as classes. No entanto, os modelos LSTM foram piores, com desempenho médio inferior a 50% para medida F1.

Figura 6.4: Diferenciadores D-AD: Performance F1 (AM vs AA)



(a) Modelos LSTM (b) Modelos CNN

6.2.3.2 Discussão

De modo geral, os classificadores diferenciadores apresentaram uma performance muito inferior, quando comparados aos modelos especialistas. Em relação à abordagem empregada para o treinamento desses modelos, tanto a técnica de ajuste fino manual, quanto automático apresentaram desempenhos similares ao longo dos experimentos. Entre as arquiteturas exploradas, os modelos CNN foram os que apresentaram menor custo computacional para treinamento, seguidos dos modelos LSTM.

A dificuldade para treinar os modelos diferenciadores, associada à baixa performance observada nos resultados, indicam que a tarefa de identificação automática de padrões para distinguir entre os transtornos alvo é ainda mais difícil do que a identificação da presença/ausência de cada condição mental. Entre os tipos de diferenciação, verifica-se que a tarefa de distinguir o transtorno depressivo da condição de comorbidade parece ser a mais difícil, uma vez que os modelos D-AD apresentaram as menores performances para identificação de cada condição alvo, segundo a medida F média.

Os modelos apresentados na Tabela 6.4 foram adotados para compor o modelo *DAC Stacking DT*. Embora apresentem performance reduzida em relação aos classificadores especialistas, esses modelos foram usados visando explorar sua contribuição para a tarefa de identificação das condições alvo, em especial da comorbidade.

6.3 Experimento #2: *Embeddings* Pré-Treinados

Esse conjunto de experimento foi projetado com o objetivo de avaliar se o uso de diferentes *embeddings* produzem maior impacto na performance dos classificadores fracos especialistas para cada condições alvo. Esses experimentos limitam-se à arquitetura LSTM e visam responder à QP4 formulada no Capítulo 1, i.e "*Existe diferença de desempenho para esta tarefa de classificação advindo do uso de word embeddings pré-treinados genéricos e/ou de domínio?*". Os *embeddings* selecionados foram produzidos a partir diferentes corpora. Com isso, espera-se promover variabilidade para modelos *DAC Stacking*, oferecendo diferentes contextos analisados pelos classificadores fracos para a tomada de decisão em relação a classificação final de uma dada instância.

6.3.1 Método

Os experimentos foram projetados para explorar o uso de *embeddings* diferentes para cada arquitetura e condição alvo. Para tanto, foram adotados os seguintes *embeddings* pré-treinados:

- *Propósito Geral*: GloVe 6B, Glove Twitter e Google News;
- *Domínio*: *All Diagnosed User* e *Target Diagnosed User* gerados a partir do corpus SMHD usando os algoritmos GloVe, Word2Vec Skip-gram e CBOW, conforme detalhado na Seção 5.4.

Os testes foram executados considerando somente a arquitetura base LSTM (Tabela 5.2). Para comparação, adotou-se como linha de base modelos que utilizam o *embedding* pré-treinado de propósito geral GloVe 6B, com aprendizado estático.

Ao todo foram desenvolvidos 54 modelos os quais são resultantes da combinação do conjunto de *embeddings* definidos para experimentação, do tipo de aprendizado explorado no treinamento (estático e não estático) e da função especialista em cada condição alvo. Cada execução foi repetida 10 vezes, e o desempenho medido em termos das métricas Precisão (P), Revocação (R) e medida F (F1). Os resultados apresentados são a média destas execuções.

Para análise de significância estatística foi usado o Teste Student T bicaudal pareado ($alpha = 0.5$). Diferentes comparativos foram realizados, visando analisar o impacto das seguintes variáveis:

- *Abordagem de Aprendizado*: comparativo de performance entre a abordagem de aprendizado estática e não estática. Como hipótese nula ($p - value >= 0.05$), adotou-se que não há diferença advindo da variação do tipo de aprendizado, considerando uma mesma condição alvo, *embedding* e arquitetura;

- *Embeddings Pré-Treinados*: para cada condição alvo, tipo de aprendizado e arquitetura, analisou-se o impacto de variar entre dois tipos de *embedding* pré-treinados. Nesse comparativo, a hipótese nula formulada é que não há diferença advinda da variação entre dois *embeddings* pré-treinados analisados ($p - value >= 0.05$).

- *Arquiteturas*: para os *embeddings* de melhor performance em cada condição alvo, realizou-se um comparativo entre as diferentes arquiteturas, onde manteve-se o mesmo tipo de aprendizado, *embedding* pré-treinado e condição alvo. Nessa avaliação, a hipótese nula é de que não existe diferença de performance entre duas arquiteturas, quando considerando uma mesma condição alvo, *embedding* pré-treinado e tipo de aprendizado ($p - value >= 0.05$).

Para cada *embedding* pré-treinado, foram estabelecidos os seguintes comparativos: (1) somente entre os *embeddings* de propósito geral; (2) somente entre os *embeddings* de domínio e (3) considerando os melhores resultados obtidos entre todos os *embeddings* experimentados (propósito geral e domínio). Esses comparativos visaram identificar se algum *embedding* produziu um ganho de performance, considerando cada condição alvo.

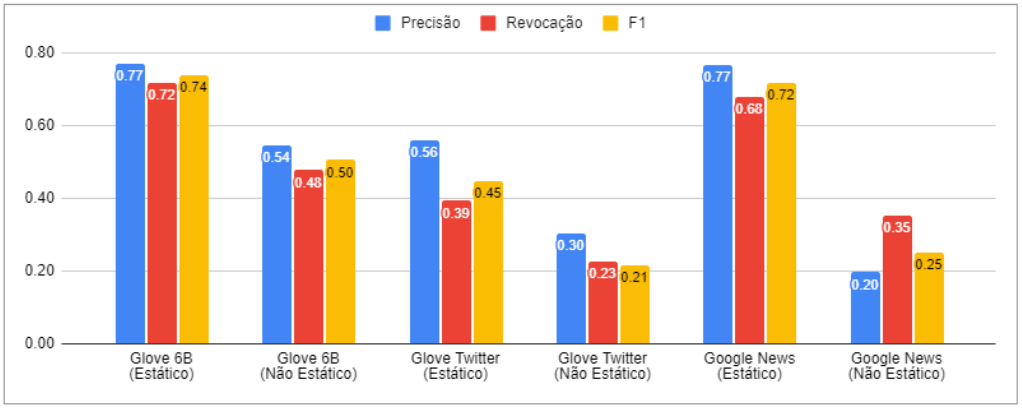
6.3.2 Resultados

6.3.2.1 Embeddings de Propósito Geral

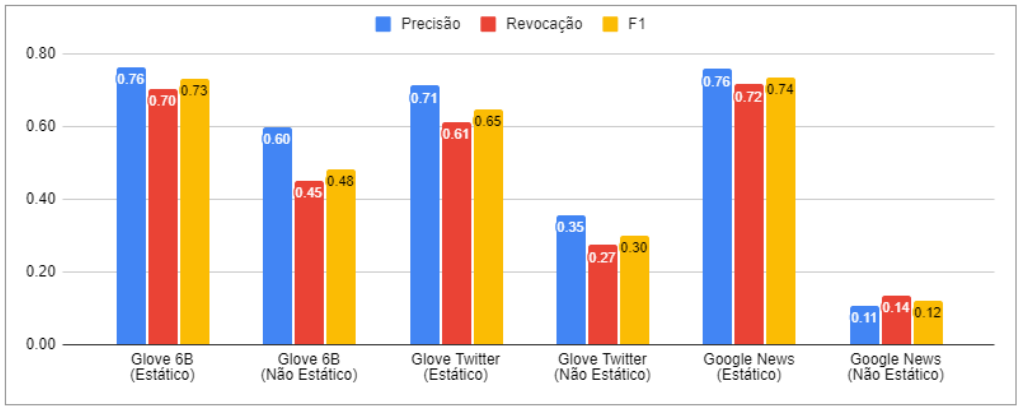
Um comparativo de performance considerando somente os *embeddings* pré-treinados de propósito geral é apresentado na Figura 6.5, destacando a média dos resultados P, R e F1 para cada condição alvo. De modo geral, a abordagem de aprendizado estático apresentou performance média melhor do que o aprendizado não estático para todos os *embeddings* de propósito geral testados e estas diferenças são estatisticamente significativas. Apenas uma exceção foi observada nos classificadores de comorbidade (Figura 6.5 (c)), onde o *embedding* GloVe Twitter com aprendizado não estático apresentou performance ligeiramente melhor (1 pp) do que seu teste com aprendizado estático.

Analisando o impacto de performance entre as abordagens de aprendizado estático

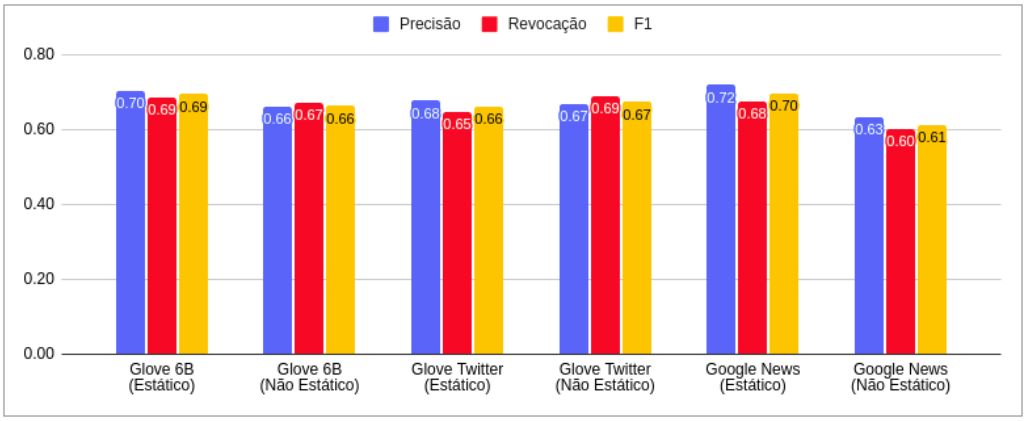
Figura 6.5: LSTM: *Embeddings* Propósito Geral (6B vs Twitter vs Google News)



(a) Classificadores de Ansiedade



(b) Classificadores de Depressão



(c) Classificadores de Comorbidade

e não estático, observam-se as seguintes diferenças estatisticamente significativas: (1) GloVe 6B, nos modelos dedicados à classificação de ansiedade e depressão; (2) Glove Twitter, nos modelos para as três condições alvo; e (3) Google News, apenas nos modelos especialistas em depressão.

Considerando a performance apresentada pelos classificadores usando *embeddings* de propósito geral, verifica-se que o GloVe 6B (estático) foi o que manteve o me-

lhor desempenho em todas as condições alvo, sendo significativamente melhor em termos de medida F para os classificadores especialistas em ansiedade e depressão. Observa-se ainda que esse modelo foi inferior em 1 pp para a comorbidade, onde o *embedding* Google News (estático) apresentou melhor performance.

6.3.2.2 Embeddings de Domínio

A Figura 6.6 apresenta um comparativo considerando a média de performance para os *embeddings* de domínio explorados, segundo cada condição alvo. De modo geral, nota-se que os modelos que usam *embeddings* gerados com Word2Vec desempenham melhor do que aqueles gerados com GloVe. Considerando a performance entre as condições alvo, verifica-se que os classificadores de comorbidade apresentaram uma distribuição de resultados similar para a maioria dos *embeddings* de domínio testados.

Em relação à abordagem de aprendizado, observa-se que para os classificadores de ansiedade e depressão, o aprendizado estático apresentou ganhos estatisticamente significativos em diferentes métricas. Para os classificadores de comorbidade, a diferença entre os tipos de aprendizado não foi significativa, à exceção do *embedding All Diagnosed User* gerado com a técnica Word2Vec (algoritmo CBOW), o qual apresentou melhor performance com aprendizado não estático para as métricas F1 e revocação.

Considerando a fonte de dados e algoritmo usado para a formação dos *embeddings*, destacam-se os seguintes resultados:

- *Target Diagnosed Users*: apresentou os melhores resultados em termos de medida F média, considerando as três condições alvo, quando gerados pela técnica Word2Vec (algoritmo Skip-gram) e abordagem estática para o treinamento dos classificadores. A segunda melhor performance para essa fonte de dados foi registrada para os classificadores de comorbidade, quando treinados com *embeddings* gerado pelo algoritmo Glove com ambas abordagens de aprendizado, seguido do Skip-gram com aprendizado não estático;

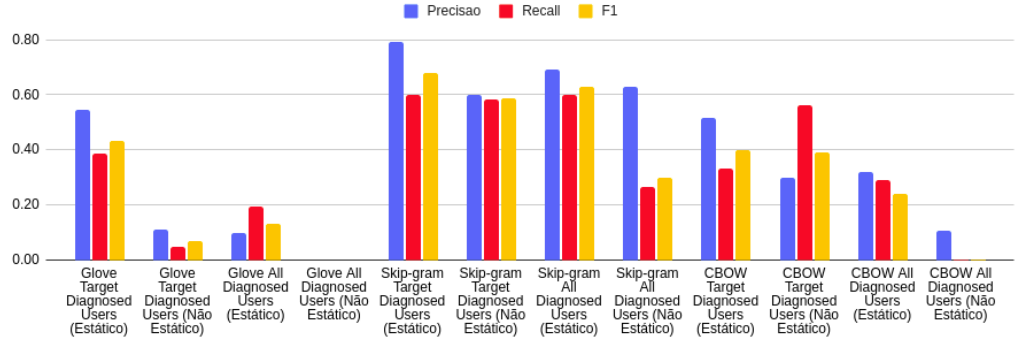
- *All Diagnosed Users*: os melhores resultados foram observados para os classificadores de comorbidade com algoritmo Word2Vec Skip-gram tanto com aprendizado estático, quanto não estático. Os classificadores de ansiedade também apresentaram resultados similares aos de comorbidade quando treinado com esse *embedding* usando abordagem estática. Já os classificadores de depressão apresentaram resultados similares aos demais, quando treinados com o *embedding* gerado pelo algo-

ritmo Word2Vec CBOW, com abordagem estática.

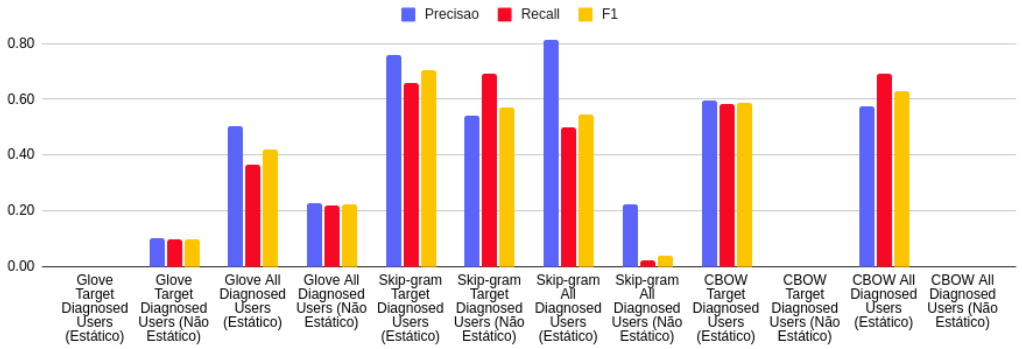
6.3.3 Discussão

A Tabela 6.7 apresenta um comparativo entre os *embeddings* pré-treinados de propósito geral e de domínio, considerando os resultados obtidos para a classificação de cada condição alvo. Os resultados estão expressos em termos das diferenças em relação ao *baseline* (GloVe 6B com aprendizado estático). As células em destaque mostram os melhores resultados para cada classificador de transtorno, usando a medida F como métrica

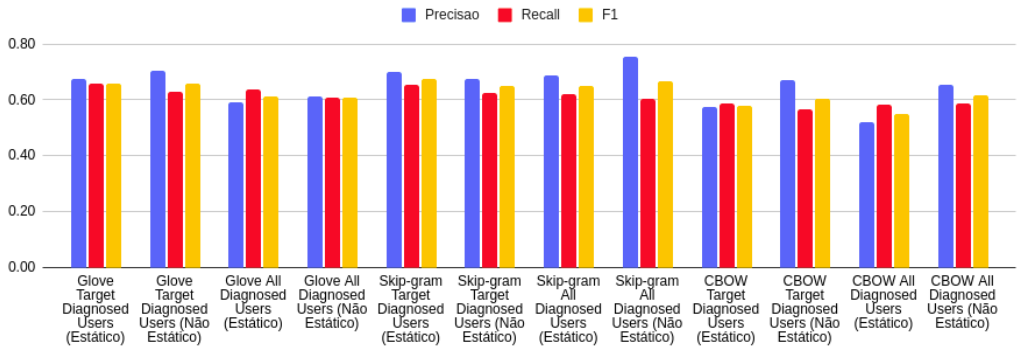
Figura 6.6: LSTM: *Embeddings* de Domínio



(a) Classificadores de Ansiedade



(b) Classificadores de Depressão



(c) Classificadores de Comorbidade

Figura 6.7: Comparativo: Performance *Embeddings* Pré-treinados GloVe 6B vs Domínio.

	Fonte de Dados	Propósito Geral (6B)			*All Diagnosed Users*									*Target Diagnosed Users*								
Modelos	WE Algoritmo	Glove			Word2Vec Skip-gram			Word2Vec CBOW			Glove			Word2Vec Skip-gram			Word2Vec CBOW			Glove		
	Aprendizado	Estático			Não-Estático			Não-Estático			Não-Estático			Estático			Estático			Estático		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
L-CA	Métricas	0,73	0,62	0,67	0,10	0,03	0,06	0,03	-0,02	0,00	0,12	0,02	0,06	-0,02	-0,01	-0,02	0,10	0,02	0,06	0,11	-0,17	-0,03
	Aprendizado	Estático			Estático			Estático			Estático			Estático			Não-Estático			Não-Estático		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
L-CD	Métricas	0,75	0,77	0,76	-0,05	0,14	0,05	0,17	0,10	0,13	0,13	0,11	0,12	0,00	0,09	0,05	0,08	-0,04	0,03	0,00	0,13	0,07
	Aprendizado	Estático			Não-Estático			Não-Estático			Não-Estático			Estático			Estático			Não-Estático		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
L-CAD	Métricas	0,72	0,58	0,65	-0,01	-0,08	-0,04	-0,05	-0,09	-0,07	-0,03	0,07	0,04	0,02	-0,05	-0,02	0,15	-0,12	0,02	0,13	-0,02	0,05

de ordenação.

Como mencionado anteriormente, o classificador de comorbidade foi o mais beneficiado pelo uso de *embeddings* de domínio. Esse resultado se mantém ao compará-lo ao Glove 6B, onde observa-se um ganho de 7 pp (pontos percentuais) na medida F, com um equilíbrio entre revocação e precisão. No entanto, o desempenho dos classificadores de depressão foi inferior em todos os casos, à exceção de um único, onde foi observado um ligeiro aumento na revocação (4 pp). Por fim, o classificador de ansiedade apresentou algumas melhorias. No caso mais significativo (3 pp na medida F), observa-se um aumento significativo na revocação em detrimento da precisão.

Esse conjunto de experimentos permitiu concluir que há poucas diferenças advindas do uso de *embeddings* pré-treinados de domínio. Contudo, embora a contribuição dos *embeddings* relacionados ao domínio seja limitada em termos de ganho de performance, ela ainda pode ser válida como um recurso para aumentar a variabilidade da solução *DAC Stacking*. Por essa razão, os *embeddings* de melhor performance em cada condição alvo foram selecionados para uma etapa de ajuste fino para treinamento desses modelos, onde alcançaram o desempenho mostrado na Tabela 6.1. Espera-se que o uso desses diferentes *embeddings* na formação dos classificadores fracos possibilite a recuperação de contextos diferentes, os quais podem auxiliar na identificação de características influentes para a classificação de cada condição mental.

6.4 Experimento #3: Avaliação Quantitativa das Variações do *DAC Stacking*

Este experimento visa responder diretamente às questões de pesquisa QP1 e QP3, a saber:

- QP1: "*Uma solução do tipo stacking ensemble é efetiva para o problema de clas-*

sificação multirrótulo envolvendo os transtornos de ansiedade, depressão e sua comorbidade?"

- QP3: "*Qual topologia, em termos de função para os classificadores fracos e tipos de arquiteturas de aprendizado profundo, tem melhor desempenho nesta tarefa de classificação?*"

O experimento também complementa a investigação relativa a QP2, quanto à confirmação das diferentes arquiteturas para prover variabilidade. O restante desta seção apresenta o método para os experimentos e discute os resultados desta avaliação quantitativa.

6.4.1 Método

Para avaliar o desempenho do modelo *DAC Stacking* e suas variações para a tarefa de classificação multirrótulo seguiu-se com a proposta descrita na Seção 5.7.1 em termos de *baselines* e métricas de desempenho. Foram gerados e avaliados os seguintes modelos:

- *Baseline L, C, H*: adaptação para o modelo multirrótulo das arquiteturas LSTM (L), CNN (C) e Híbrida (H) das configurações base dos respectivos classificadores fracos, conforme detalhado na Seção 5.7.1. As parametrizações de cada configuração base estão detalhadas nas Tabelas 5.2, 5.3 e 5.4, respectivamente;

- *DAC Stacking*: vários modelos foram desenvolvidos para a solução *DAC Stacking* base e suas variações *EC* e *DT*, através de experimentação de distintas topologias para o *Nível 0*. Foram gerados modelos com topologia homogêneas (de mesma arquitetura) e heterogêneas, usando os melhores classificadores fracos de cada arquitetura/condição alvo resultantes do Experimento #1 (Tabelas 6.1, 6.2, 6.3). As Tabelas 6.5, 6.6 e 6.7 listam os modelos gerados para *DAC Stacking*, *DAC Stacking EC* e *DAC Stacking DT*, respectivamente, junto com os classificadores fracos que compõem cada topologia. Todos estes modelos foram treinados e testados usando o conjunto de dados SMHD A-D-AD, segundo método descrito na Seção 5.6.3. Além disto, os modelos das variações *EC* e *DT* são precedidos também por estes prefixos. Como na Seção 6.3, foram usados os prefixos *L*, *C* e *H* para denotar os modelos de topologia homogênea (arquitetura LSTM, CNN e Híbrida, respectivamente). Para

Tabela 6.5: Topologia do *Nível 0* dos Modelos *DAC Stacking*

Modelo	Arquitetura	Classificador Fracos Especilistas	
		Ansiedade	Depressão
L	LSTM	L-CA1, L-CA2, L-CA3	L-CD1, L-CD2, L-CD3
C	CNN	C-CA1, C-CA2, C-CA3	C-CD1, C-CD2, C-CD3
H	Híbrida	H-CA1, H-CA2, H-CA3	H-CD1, H-CD2, H-CD3
LCH	LSTM, CNN, Híbrida	L-CA2, C-CA3, H-CA1	L-CD2, C-CD2, H-CD2

Tabela 6.6: Topologia do *Nível 0* dos Modelos *DAC Stacking EC*

Modelo	Arquitetura	Classificador Fracos Especialistas		
		Ansiedade	Depressão	Comorbidade
EC-L	LSTM	L-CA1, L-CA2, L-CA3	L-CD1, L-CD2, L-CD3	L-CAD1, L-CAD2, L-CAD3
EC-C	CNN	C-CA1, C-CA2, C-CA3	C-CD1, C-CD2, C-CD3	C-CAD1, C-CAD2, C-CAD3
EC-H	Híbrida	H-CA1, H-CA2, H-CA3	H-CD1, H-CD2, H-CD3	H-CAD1, H-CAD2, H-CAD3
EC-LCH	LSTM, CNN, Híbrida	L-CA2, C-CA3, H-CA1	L-CD2, C-CD2, H-CD2	L-CAD3, C-CAD1, H-CAD2

Tabela 6.7: Topologia do *Nível 0* dos Modelos *DAC Stacking DT*

Modelo	Arquitetura Especialistas - Diferenciadores	Classificador Fracos		
		Especialistas		Diferenciadores
		Ansiedade	Depressão	A-D, A-AD e/ou D-AD
DT-L	LSTM	L-CA1, L-CA2, L-CA3	L-CD1, L-CD2, L-CD3	L-CA-D1, L-CA-AD1, L-CA-AD2
DT-C	CNN	C-CA1, C-CA2, C-CA3	C-CD1, C-CD2, C-CD3	C-CA-D1, C-CA-AD1, C-CD-AD1
DT-LCH-LC	LSTM, CNN, Híbrida - LSTM, CNN	H-CA1, H-CA2, H-CA3	H-CD1, H-CD2, H-CD3	C-CA-D1, L-CA-AD1, C-CD-AD1
DT-LCH-C	LSTM, CNN, Híbrida - CNN	L-CA2, C-CA3, H-CA1	L-CD2, C-CD2, H-CD2	C-CA-D1, C-CA-AD1, C-CD-AD1

as topologias heterogêneas, os prefixos são formados por combinações destas letras. Por exemplo, os modelos formados pela combinação das três arquiteturas são denominados *LCH*.

O desempenho de cada modelo foi mensurado usando as métricas de avaliação para classificadores multirrótulo Exact Match Ratio (EMR) e Hamming Loss (HL). Para avaliar o comportamento quanto à cada classe, foram calculadas as métricas Precisão (P), Revocação (R) e medida F (F1) adaptadas para classificadores multirrótulo (Seção 2.6.2).

Os resultados reportados correspondem à performance média de cada modelo sobre o conjunto de testes SMHD A-D-AD. Essa média foi obtida pelo número de repetições resultante do processo de treinamento considerando a validação cruzada com *k-fold* = 5. Para verificação de significância estatística dos resultados foi usado o Teste Student T bicaudal pareado ($alpha = 0.5$). Como hipótese nula ($p - value >= 0.05$), adota-se que não há diferença de performance entre dois modelos *DAC Stacking* comparados, considerando a variação proposta para composição do *Nível 0*. Os resultados dos testes estatísticos são detalhados no Apêndice E, juntamente com a relação de valores de média e desvio padrão para cada experimento.

Finalmente, para o classificador de melhor performance média para o modelo *DAC*

Stacking e suas variações (*EC* e *DT*), realizou-se uma análise dos tipos de erro mais frequentes, considerando o desempenho de classificação sobre todo o conjunto de amostras de teste. Para definir o modelo de melhor performance adotou-se o critério melhores resultados para as métricas EMR e HL, associado a uma performance equilibrada em termos de medida F para as classes representando as condições alvo.

6.4.2 Resultados

6.4.2.1 DAC Stacking

A Tabela 6.8 apresenta o desempenho médio do experimento para cada topologia para o *Nível 0* avaliado para os modelos *DAC Stacking* (Tabela 6.5), juntamente com os modelos *baseline*. Em relação aos *baselines*, todos os modelos *DAC Stacking* apresentaram performance estatisticamente superior. Para a métrica EMR a diferença de performance variou entre 7 a 15 pp, enquanto que para a métrica HL, a diferença foi de 3 a 8 pp. As métricas por classe também foram estatisticamente superiores.

Comparando o desempenho dos modelos *DAC Stacking* homogêneos em relação aos respectivos *baselines*, observa-se que os maiores ganhos de performance foram alcançados pela arquitetura CNN (*C*) para a métrica EMR, e pela arquitetura LSTM (*L*) para a métrica HL. O comparativo entre o modelo *DAC Stacking* heterogêneo (*LCH*) e cada modelo linha de base, mostra que a maior diferença de performance para as métricas EMR (13 pp) e HL (7 pp) foram registradas em relação aos modelos *Baseline C* e *Baseline L*, respectivamente.

Entre as topologias *DAC Stacking* avaliadas, o modelo *L* foi o que apresentou o melhor desempenho considerando a métrica estrita EMR (4 a 6 pp). Contudo, observa-se que este comportamento é justificado principalmente pelo desempenho desta topologia na classe Controle, com resultados estatisticamente superiores em termos de revocação e

Tabela 6.8: *DAC Stacking*: Performance Média

Modelo	**Multi Tarefa**		**Controle (C)**			**Ansiedade (A)**			**Depressão (D)**			**Comorbidade (AD)**		
	EMR	**HL**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
L	0,46	0,29	0,67	0,77	0,72	0,62	0,69	0,65	0,57	0,66	0,61	0,33	0,72	0,45
C	0,42	0,28	0,74	0,63	0,68	0,62	0,73	0,67	0,58	0,72	0,64	0,33	0,75	0,46
H	0,42	0,29	0,75	0,64	0,69	0,60	0,74	0,67	0,57	0,67	0,61	0,33	0,76	0,46
LCH	0,40	0,30	0,75	0,59	0,66	0,58	0,81	0,67	0,56	0,71	0,63	0,32	0,79	0,46
Baseline L	0,33	0,37	0,57	0,63	0,60	0,57	0,39	0,45	0,55	0,35	0,41	0,29	0,37	0,32
Baseline C	0,28	0,33	0,73	0,48	0,53	0,62	0,36	0,45	0,62	0,37	0,46	0,33	0,36	0,34
Baseline H	0,31	0,36	0,60	0,52	0,54	0,54	0,47	0,49	0,55	0,46	0,50	0,30	0,44	0,35

medida F (à exceção de *H*). A medida F média de *L* é superior em 3 a 6 pp, comparada às outras topologias. Contudo, este modelo possui resultados inferiores para as medidas F das condições alvo, principal interesse deste trabalho.

Em termos da métrica mais flexível HL, as topologias apresentaram resultados estatisticamente comparáveis, com exceção da diferença 1 pp apresentada entre os modelos *L* e *C*. Avaliando a performance da medida F para as condições alvo, verifica-se que o modelo *C* foi o que apresentou o melhor desempenho, sendo significativamente superior ao modelo *L* em 2 pp para as classes Ansiedade e Depressão, e 1 pp para Comorbidade. Comparado ao modelo *H*, *C* apresenta uma perda de 2 pp para a classe Controle, mas ganho de performance de 1 pp para Ansiedade e 2 pp para Depressão. O desempenho desses modelos são comparáveis para a Comorbidade. Em relação ao modelo *LCH*, *C* apresentou performance superior em 2 pp para a classe Controle e 1 pp para a classe Ansiedade, sendo esses modelos comparáveis para a classe Comorbidade. Por fim, o comparativo entre *H* e *LCH* mostra que *H* é superior para a classe Controle em 3 pp, inferior em 1 pp para Depressão e Ansiedade, e comparável para Comorbidade.

Estes resultados indicam que, no geral, os modelos *C* e *LCH* apresentaram desempenhos similares, sendo *C* ligeiramente superior para as métricas relacionadas às condições alvo. Por essa razão, considerou-se que a topologia representada por *C* é a melhor para o *Nível 0* do *DAC Stacking*.

Para este modelo, procedeu-se então a análise dos tipos de erros mais frequentes em termos de diferenciação entre (1) usuários saudáveis e diagnosticados e (2) condições mentais distintas.

Em relação à diferenciação entre usuários Controle e diagnosticados, verificou-se que do total de amostras de teste, 7% apresentaram erro para as situações predizer usuário saudável como diagnosticado, sendo estes erros distribuídos em 2% para Depressão e 5% para a Comorbidade. Em relação ao erro inverso de predizer um usuário diagnosticado como saudável, verifica-se uma taxa de 9% no total de amostras de teste, distribuída em 3% para Ansiedade, 4% para Depressão e 2% para a Comorbidade.

Considerando os erros entre os diferentes tipos de usuários diagnosticados, a análise revelou que o erro mais comum é predizer a condição de Comorbidade para usuários diagnosticados com somente Ansiedade (13%) ou somente Depressão (12%), o que explica em boa parte a baixa precisão. Também foram observados alguns erros envolvendo os usuários diagnosticados com (1) Ansiedade, dos quais 2% foram erroneamente classificados como depressivos e (2) Comorbidade, onde 1% foi classificado como depressivo.

Não foram observados casos de erro envolvendo usuários diagnosticados com Depressão rotulados como ansiosos.

Nota-se que o ponto fraco de todos as topologias *DAC Stacking* é o desempenho em relação à Comorbidade, devido à baixa precisão. As variações *DAC Stacking EC* e *DAC Stacking DT* visam verificar se a inclusão de classificadores especialistas para a Comorbidade ou diferenciadores entre as condições contribuem à melhoria de desempenho da solução *DAC Stacking*.

6.4.2.2 DAC Stacking EC

O *DAC Stacking EC* foi proposto com o intuito de verificar se a inclusão de classificadores especialistas em Comorbidade no comitê melhoraria o desempenho do modelo. A Tabela 6.9 apresenta a média de resultados das execuções para as diferentes topologias proposta de *DAC Stacking EC* (Tabela 6.6), junto aos *baselines*. Novamente, todos os modelos *DAC Stacking EC* superaram os *baselines*, sendo o desempenho médio superior para as métricas EMR (entre 7 a 15 pp), e HL (4 a 8 pp). Os resultados para estas métricas são similares àqueles alcançados pelas topologias *DAC Stacking*, não havendo diferença estatística.

Tabela 6.9: *DAC Stacking EC*: Performance Média

Modelo	**Multi Tarefa**		**Controle (C)**			**Ansiedade (A)**			**Depressão (D)**			**Comorbidade (AD)**		
	EMR	**HL**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
EC-L	0,46	0,29	0,67	0,78	0,72	0,61	0,69	0,65	0,57	0,64	0,60	0,32	0,71	0,45
EC-C	0,42	0,28	0,74	0,66	0,69	0,63	0,68	0,65	0,58	0,71	0,64	0,33	0,72	0,45
EC-H	0,43	0,29	0,72	0,68	0,70	0,60	0,76	0,67	0,56	0,69	0,62	0,32	0,76	0,44
EC-LCH	0,41	0,29	0,75	0,61	0,67	0,62	0,73	0,66	0,57	0,69	0,62	0,33	0,75	0,46
Baseline L	0,33	0,37	0,57	0,63	0,60	0,57	0,39	0,45	0,55	0,35	0,41	0,29	0,37	0,32
Baseline C	0,28	0,33	0,73	0,48	0,53	0,62	0,36	0,45	0,62	0,37	0,46	0,33	0,36	0,34
Baseline H	0,31	0,36	0,60	0,52	0,54	0,54	0,47	0,49	0,55	0,46	0,50	0,30	0,44	0,35

Comparando o desempenho das diferentes topologias *DAC Stacking EC*, verifica-se novamente que o modelo baseado em LSTM (*EC-L*), foi significativamente superior em relação à métrica estrita EMR. Em relação à metrica HL, com exceção da topologia *EC-L* que significativamente superior em 1 pp em relação à arquitetura *EC-C*, todas as demais topologias são estatisticamente comparáveis.

Em relação às classes alvo, cada variação topológica do *DAC Stacking EC* foi em média superior em uma delas, mas praticamente não há significância estatística entre as diferenças. Considerando a medida F para cada classe alvo, tem-se que:

- *Classe Controle*: *EC-L* apresentou o melhor resultado, mas que é significativamente

superior somente a *EC-LCH*;

- *Classe Ansiedade*: a melhor performance média foi obtida por *EC-H*, mas que é significativamente superior somente a *EC-L*;

- *Classe Depressão*: *EC-C* apresentou o maior resultado médio, mas que é significativamente superior a *EC-L* (4 pp) e *EC-H* (2 pp);

- *Classe Comorbidade*: a melhor performance média foi observada para *EC-LCH*, mas esta superioridade não é estatisticamente significativa em relação a nenhuma outra topologia.

Dado que nenhuma topologia do *DAC Stacking EC* é claramente superior, procedeu-se a avaliação de erros em todos os modelos, verificando-se os seguintes comportamentos:

- *Predizer usuário saudável como diagnosticado*: a taxa de erro variou entre 6 a 9%, sendo a menor taxa registrada para o modelo *EC-C*, distribuída em 2% para Depressão e 5% para Comorbidade. Não foram registrados casos envolvendo a Ansiedade;

- *Predizer usuário diagnosticado como saudável*: a taxa de erro variou entre 8 e 15%, sendo a menor taxa observada no modelo *EC-LCH*. Em *EC-LCH*, esse erro está distribuído em 3% para Ansiedade, 3% para Depressão e 2% para Comorbidade;

- *Erros para a predição das condições alvo*: novamente, o erro mais comum é predizer condição de Comorbidade para os usuários diagnosticados com Ansiedade (13%) ou com Depressão (12%). Com relação a predizer usuários depressivos como ansiosos, destaca-se que não foram observados erros para os modelos *EC-C* e *EC-LCH*. O inverso é observado para os modelos *EC-L* e *EC-H*, os quais não registraram erros de predição envolvendo usuários ansioso como depressivo.

Por fim, um comparativo de performance foi realizado entre o modelo *DAC Stacking C* e todas as varições de topologia propostas para *DAC Stacking EC*. Dessa análise, constatou-se que o *DAC Stacking C* é comparável a todos os modelos em termos de EMR e HL, com exceção do modelo *DAC Stacking EC-L*. Este modelo foi significativamente melhor que o *DAC Stacking C* para as métricas EMR e HL, devido a performance apresentada para a classe Controle. No entanto, analisando a performance para as condições

alvo entre esses modelos, verificou-se que *DAC Stacking C* foi significativamente superior ao modelo *DAC Stacking EC-L* em todas as classes.

Com estes resultados, pode-se afirmar que em nenhuma das topologias experimentadas a inclusão dos classificadores especialistas em Comorbidade promoveu ganho significativo de performance, quando comparada com *DAC Stacking C*. Além disso, nenhuma dentre elas resolveu o problema da baixa precisão para a Comorbidade anteriormente identificado no *DAC Stacking C*.

6.4.2.3 DAC Stacking DT

Esta última avaliação tem por objetivo verificar se a inclusão dos classificadores diferenciadores auxilia no melhor desempenho do modelo *DAC Stacking*, em particular em relação aos erros relativos à Comorbidade. A Tabela 6.10 apresenta a média de resultados das execuções para as diferentes topologias avaliadas no *DAC Stacking DT* (Tabela 6.7). Cabe lembrar que não há classificadores diferenciadores seguindo a arquitetura híbrida, e portanto as topologias heterogêneas produzidas referem-se à inclusão de classificadores fracos diferenciadores formados somente pela arquitetura CNN (*DT-LCH-C*), e formados pelas arquiteturas LSTM e CNN (*DT-LCH-LC*).

Tabela 6.10: *DAC Stacking DT*: Performance Média

Modelo	**Multi Tarefa**		**Controle (C)**			**Ansiedade (A)**			**Depressão (D)**			**Comorbidade (AD)**		
	EMR	**HL**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
DT-L	0,46	0,29	0,68	0,77	0,72	0,61	0,71	0,66	0,57	0,66	0,61	0,32	0,73	0,45
DT-C	0,42	0,28	0,73	0,65	0,69	0,62	0,72	0,67	0,57	0,72	0,63	0,33	0,75	0,46
DT-LCH-LC	0,42	0,29	0,74	0,62	0,67	0,60	0,76	0,66	0,58	0,69	0,62	0,32	0,69	0,43
DT-LCH-C	0,40	0,30	0,74	0,60	0,66	0,59	0,76	0,66	0,55	0,76	0,64	0,31	0,77	0,44
Baseline L	0,33	0,37	0,57	0,63	0,60	0,57	0,39	0,45	0,55	0,35	0,41	0,29	0,37	0,32
Baseline C	0,28	0,33	0,73	0,48	0,53	0,62	0,36	0,45	0,62	0,37	0,46	0,33	0,36	0,34
Baseline H	0,31	0,36	0,60	0,52	0,54	0,54	0,47	0,49	0,55	0,46	0,50	0,30	0,44	0,35

Novamente observa-se que todas variações de *DAC Stacking DT* apresentaram desempenho superior aos *baselines*, sendo a diferença de 7 a 15 pp para a métrica EMR, e de 5 a 9 pp para a métrica HL. Contudo, estes valores não diferem daqueles já obtidos pelos modelos DAC Stacking e DAC Stacking EC.

A topologia formada somente pela arquitetura LSTM (DT-L), mais uma vez, apresentou a melhor performance para a métrica estrita EMR, sendo significativamente superior em 4 pp quando comparado aos modelos *DT-C* e *DT-LCH-LC*, e 6 pp quando comparada ao *DT-LCH-C*. Assim como em *DAC Stacking L*, este comportamento é bem explicado pelo bom desempenho na classe Controle. Em termos de medida F, o modelo

DT-L apresentou performance significativamente superior aos demais modelos para esta classe. Em termos de HL, o melhor resultado foi observado no modelo *DT-C*, sendo este resultado significativamente superior apenas ao *DT-LCH-LC*.

Considerando as condições alvo, embora o desempenho médio dos modelos *DT-C*, *DT-LCH-LC* e *DT-LCH-C* para as condições alvo sejam distintos, são estatisticamente comparáveis para as classes Ansiedade e Depressão. Para a classe Comorbidade, o modelo *DT-C* apresentou a maior performance média, porém esse resultado é significativamente superior somente quando comparado ao modelo *DT-LCH-LC* (3 pp).

A análise de erros para os quatro modelos permitiu verificar que os modelos *DT-C* e *DT-LCH-C* apresentaram as menores taxas para os erros envolvendo a classificação entre os usuários saudáveis e diagnosticados, sem diferença significativa entre esses modelos. O mesmo foi observado para os erros envolvendo a identificação das condições alvo entre os usuários diagnosticados. Para ambos os modelos, o erro de predizer um usuário saudável como diagnosticado foi de 8% distribuído, entre Depressão (1%) e Comorbidade (7%). O erro de predizer usuário diagnosticado como saudável também ocorre em 8% dos casos, sendo distribuído em Ansiedade (3%), Depressão (3%) e Comorbidade (2%).

Quanto à análise de erros entre as condições alvo, verifica-se novamente que o erro mais comum é predizer a condição de Comorbidade para usuários somente ansiosos (13%) ou somente depressivos (12%). Em menor frequência, foram observados erros envolvendo usuários depressivos classificados incorretamente como ansiosos ou com condição de Comorbidade. Esses erros somam 2% do total de amostras de teste analisadas, sendo igualmente distribuídos. Não foram registrados erros envolvendo a classificação de usuários diagnosticados com Comorbidade como ansiosos.

6.4.3 Discussão

Considerando as topologias homogêneas, de um modo geral os experimentos mostraram que existe uma diferença significativa de performance entre os modelos compostos apenas pela arquitetura LSTM quando comparados àqueles baseados em CNN ou Híbridos. Contudo, a arquitetura LSTM contribuiu mais para a identificação da classe Controle, enquanto as outras arquiteturas apresentaram melhor performance para identificação das condições mentais. Entre as arquiteturas CNN e Híbrida, poucas diferenças significativas foram notadas. Por exemplo, considerando o *DAC Stacking EC*, apenas a medida F para Depressão apresentou diferença significativa a favor da arquitetura CNN. Considerando o

custo computacional para treinar esses modelos, pode-se afirmar que as arquiteturas CNN são mais vantajosas, pois apresentaram performances similares a um custo menor. Mas cabe investigar se a restrição a uma única arquitetura de aprendizado profundo resulta na variabilidade necessária a um classificador do tipo comitê.

Com relação à inclusão de classificadores fracos especialistas na condição de Comorbidade proposta para a variação *DAC Stacking EC*, conclui-se que sua presença não promoveu ganho de performance. Mais ainda, a presença desse especialista não resultou ganhos significativos de performance para a própria condição de Comorbidade.

A inclusão desses classificadores diferenciadores proposta para a variação *DAC Stacking DT* também não melhorou os problemas já identificados. Contudo, é de observar que o desempenho muito inferior dos classificadores diferenciadores não afetou negativamente o desempenho do comitê base.

Assim, efetuou-se um comparativo entre os modelos *DAC Stacking C*, *DAC Stacking DT-C* e *DAC Stacking DT-LCH-C*, os quais apresentaram os melhores resultados considerando todos os experimentos para a solução *DAC Stacking* proposta. De modo geral, foi possível verificar que os modelos apresentaram performance média muito similares em todas as classes. Entretanto, o modelo *DAC Stacking DT-LCH-C* apresentou uma redução de 1 pp para a taxa de erro em predizer um usuário diagnosticado como saudável, a qual é estatisticamente significativa. Por este critério, considerou-se esta topologia para a análise qualitativa do *DAC Stacking* desenvolvida na próxima seção.

Além deste critério, pesou nesta decisão a experiência prévia de análise feita sobre um *DAC Stacking* homogêneo de arquitetura LSTM (SOUZA; NOBRE; BECKER, 2020), que mostrou que mesmo usando os *embeddings* diferentes, o conjunto de *features* apresentado era muito similar para a tomada de decisão de classificação pelo *meta-learner*. Assumindo que diferentes arquiteturas podem contribuir para a variabilidade do modelo *ensemble*, selecionou-se a arquitetura *DAC Stacking DT-LCH-C* para uma avaliação qualitativa, onde analisa-se a relação entre as *features* relevantes para a classificação e os sintomas de cada condição alvo.

6.5 Experimento #4: Avaliação Qualitativa para Compreensão dos Padrões de Classificação

A avaliação qualitativa visou responder a questão de pesquisa QP5, i.e "*As features relevantes para a classificação de cada condição alvo representam características dos*

sintomas típicos de cada transtorno mental?". Essa análise também complementa as questões QP2 e QP4, quanto aos fatores que mais contribuem à variabilidade do modelo *DAC Stacking* (arquiteturas e/ou *embeddings*). O restante desta seção apresenta o método usado para esta avaliação e discute os resultados encontrados.

6.5.1 Método

A avaliação qualitativa seguiu o método proposto na Seção 5.7.2. Cabe lembrar que os termos foram analisados com relação a cada classificador fraco que compõe o *Nível 0*, já que a biblioteca SHAP não suporta a análise de modelos customizados.

A avaliação foi realizada considerando um subconjunto de amostras do conjunto de teste SMHD A-D-AD, previstos de acordo com o modelo *DAC Stacking DT-LCH-C*. Esse subconjunto é formado por amostras correta e incorretamente classificadas em cada classe, selecionadas de modo aleatório.

Para compreender o que os termos relevantes à classificação poderiam significar em termos de sintomas ou emoções relacionados às condições, adotou-se os seguintes critérios:

a) *Amostras corretamente classificadas*: foram analisados os termos mais frequentes considerados como influentes em pelo menos dois modelos, os quais foram selecionados conforme a condição alvo da amostra. Esses termos são pesquisados nos Dicionários de Sintomas (DS), segundo a função de busca definida na Seção 5.7.2. Por exemplo, para amostras de Ansiedade, a análise foi realizada com base nos termos influentes nos classificadores especialistas em ansiedade e nos classificadores diferenciadores A-D e A-AD. Para as amostras de Controle e Comorbidade, foram consideradas a união dos termos encontrados por cada conjunto de modelos especialistas (Ansiedade - CA(i), Depressão CD(i)), juntamente com os modelos diferenciadores;

b) *Amostras com erro*: foram analisadas conforme o conjunto de termos mais frequentes considerando todos os classificadores para compreender os erros mais comuns apontados pela análise quantitativa do modelo *DAC Stacking DT-LCH-C*: (1) classificar usuários com somente uma condição alvo como em condição de comorbidade; e (2) também foram analisados erros envolvendo usuários saudáveis e diagnosticados. Os erros foram analisados tanto em relação aos termos influentes dos classifi-

cadores fracos, quando à consolidação realizada pelo *meta-learner*.

Por fim, para avaliar se algum recurso contribuiu mais à variabilidade do comitê, i.e. variações de arquiteturas e/ou de *embeddings*, realizou-se uma análise comparativa dos termos influentes no classificadores fracos, considerando tanto o modelo *DAC Stacking DT-LCH-C*, quanto o *DT-C*. A análise foi baseada no subconjunto de amostras de teste corretamente classificadas. A Tabela 6.11 apresenta os classificadores fracos selecionados para essa análise, a qual foi estruturada como segue:

a) *Variação Arquitetural (QP2)*: para essa análise adotou-se uma arquitetura heterogênea, formada por modelos que usam o mesmo *embedding* pré-treinado (GloVe 6B);

b) *Variação de embeddings pré-treinados (QP4)*: essa análise considerou uma arquitetura homogênea, formada por modelos CNN que usam diferentes *embedding* pré-treinados.

Tabela 6.11: Variação de Recursos: Modelos selecionados para análise comparativa.

Variação	***Dac Stacking***	**Modelo**	***Embedding Pré-treinado***
Arquitetural	DT-LCH-C	L-CA2	GloVe 6B
		H-CA1	
		L-CD2	
		H-CD2	
	DT-C	C-CA1	
		C-CD1	
Embeddings Pré-treinados	DT-C	C-CA1	GloVe 6B
		C-CA2	GloVe Twitter
		C-CA3	All Diagnosed Users*
		C-CD1	GloVe 6B
		C-CD2	GloVe Twitter
		C-CD3	Target Diagnosed Users*

* Gerados usando a técnica Word2Vec Skip-gram.

6.5.2 Resultados: Amostras Classificadas Corretamente

A Tabela 6.12 apresenta os 20 termos SHAP considerados mais influentes na classificação de cada rótulo para as amostras analisadas, especificando para cada condição alvo os classificadores fracos usados na geração dos termos. A tabela também elenca quais destes termos foram encontrados em cada Dicionário de Sintomas. Nota-se que

todos os classificadores fracos, de um modo geral, estão relacionados a um bom número de termos associados ao DS Ansiedade, quer para usuários saudáveis e diagnosticados. Observa-se também que os classificadores especialistas e diferenciadores contribuíram com critérios diferentes para a tomada de decisão, em relação à classificação de ambas condições alvo. A seguir, são analisados os termos influentes em relação a cada rótulo, separados por classificadores especialistas e diferenciadores.

Tabela 6.12: Lista de termos relevantes SHAP em amostras corretamente classificadas

Condição Mental	**Os 20 Termos SHAP mais Relevantes**	**Modelo**	**Termos Relevantes SHAP encontrados nos Dicionários de Sintomas (conforme cada classificador base)**		
			Termos Comuns aos Dicionários de Ansiedade e Depressão	**Somente no Dicionário de Ansiedade**	**Somente no Dicionário de Depressão**
Ansiedade	my, will, working, everyday, me, cards, something, interesting, states, boss, paying, if shift, management, fact, can, chaos, know, accepting, last	L-CA2 C-CA3 H-CA1	my, something	chaos, if, will, cards, me, management	shift
		C-CA-D1	my	will, something, know, management, can	paying, shift, if
		C-CA-AD1	my	know, will, something	could
Depressão	things, hope, attempts, me, because, failed, feel, motivation, keep, loose, hope, if, wrong, cause, can, something, help, stuck, bad, think, always	L-CD2 C-CD2 H-CD2	cause, failed, feel, things, because, something, think	attempts, help, hope, if, me	bad, always, anything, wrong, stuck
		C-CA-D1	cause, failed, feel, keep, things	help, something, hope, can	if, bad, because, stuck
		C-CD-AD1	because, failed, feel, something, things	help, if, attempts, me, hope	bad, anything, always
Comorbidade	my, mind, night, will, ideas, discussion, afraid, experience, cold, weird, consider, obviously, really, trying, me, anymore, because, migraines, terrible, crying	L-CA2 C-CA3 H-CA1	ideas, mind, my, night	weird, really, cold, will, trying, afraid	obviously, much, experience, terrible
		L-CD2 C-CD2 H-CD2	ideas, mind, my, night, afraid, crying	cold, trying, weird, will, afraid	experience, obviously, much, consider
		C-CA-D1	afraid, ideas, migraines, mind, my	weird, cold, will, effects, anymore, can	consider, because, obviously
		C-CA-AD1	afraid, ideas, migraines, mind, my	weird, cold, will, anymore, can	because, obviously
		C-CD-AD1	because, ideas, migraines, mind, my, night	weird, really, cold, will, having, me, afraid	obviously, experience, much
Controle	why, sorry, guys, if, think, something, could, game, instead, know, calm, down, wonder, me, strange, though, mistake, maybe, same, always	L-CA2 C-CA3 H-CA1	something, think, what	if, strange, going, down, know, guys, could, why, me, instead	calm, sorry
		L-CD2 C-CD2 H-CD2	maybe, something, think	if, strange, going, down, know, guys, do, could, why, me, instead, new	calm, sorry, though, wonder
		C-CA-D1	what	new, know, something	if, sorry, could, instead
		C-CA-AD1		something, down, know, new	if, sorry, could, calm, instead, wonder
		C-CD-AD1	maybe, something, think, what	if, strange, going, know, down, guys, could, why, me, instead	calm, sorry

a) Ansiedade. Para as amostras de Ansiedade corretamente classificadas, nota-se que além de apresentar mais termos influentes pertencentes ao DS Ansiedade, esses termos indicam uma relação com alguns critérios que definem a presença desse transtorno. Conforme a função dos classificadores fracos, observou-se:

- *Especialistas*: considerando os termos influentes associados aos classificadores especialistas CA_i, nota-se que sua relação com termos próximos encontrados nos DSs podem dar subsídios para interpretação do seu significado. Por exemplo, o termo influente "chaos" é próximo aos termos "panic", "fear", "anxiety" e "confusion", que podem ser associados a um estado de ansiedade extrema, tais como episódios de ataque de pânico recentes ou iminentes. Já o termo "if" aparece associado a

"future", "possible", "could" e "would" indicando preocupação com situações futuras e dúvidas. Ainda em relação aos classificadores especialistas, observa-se a presença de poucos termos no DS Depressão. Entre eles pode-se destacar "shift" que está associado a "change" e "need" no dicionário de Depressão, e a "failed" no DS Ansiedade, esses termos sugerem a necessidade de mudança e preocupação em falhar, sentimentos que podem estar presentes também em ansiosos;

- *Diferenciadores*: verificou-se que os modelos CA-D e CA-AD acertaram a definição da classe Ansiedade. Além dos termos identificados pelos classificadores de ansiedade, esses modelos também apresentam termos comuns, tais como "know" e "something" que aparecem próximos a "thinking", "feel", "doubt", sentimentos presentes no DS Ansiedade. Especificamente o diferenciador A-D apresentou adicionalmente o termo "can" e "management" que estão associados a "help", "control" e "pressure", que poderiam ser indicativos de preocupação excessiva acerca de situações futuras, um comportamento típico dos Transtornos de Ansiedade.

b) Depressão. As amostras de Depressão corretamente classificadas apresentaram termos relevantes que estão presentes tanto no DS Depressão, quanto na relação de termos comuns entre os dois DSs. Em relação às funções dos classificadores fracos, tem-se:

- *Especialistas*: inspecionando os termos mais frequentes para Depressão, destacados pelos classificadores CD_i, encontrou-se o termo "failed" associado com "feel", "inability", "insufficient", o que poderia indicar um sentimento de incapacidade para realizar tarefas. Outro exemplo é "bad", que aparece associado a "worse", "horrible", "terrible" e "suffer", possivelmente indicando sentimentos de culpa excessiva, típicos da Depressão;

- *Diferenciadores*: entre os classificadores diferenciadores, o modelo A-D apresentou adicionalmente outros termos que poderiam reforçar os sentimentos típicos da Depressão, tais como o termo "stuck" que aparece associado a "empty" e "broken", que podem remeter ao sentimento de vazio e desesperança.

c) Comorbidade. Para a comorbidade, verificou-se uma quantidade de termos comuns entre Depressão e Ansiedade em média maiores que usuários ansiosos e saudáveis, mas inferior aos usuários depressivos. Analisando os termos mais relevantes para a comorbidade, segundo a função dos classificadores fracos, tem-se:

- *Especialistas*: verifivou-se termos incluídos somente no DS Ansiedade, como “weird” e “trying”, os quais são associados a “trouble”, “worry” e “danger”. Estes termos denotam sentimentos que podem indicar excessiva preocupação com perigo, fortemente presentes para os Transtornos de Ansiedade (e.g. Transtorno de Ansiedade Generalizada). Outros termos estão presentes somente no DS Depressão, tais como “terrible”, que aparece associado a “sad”, “melancholy” e “lonely”, sentimentos presentes em deprimidos que referem-se ao humor triste, vazio e sem esperança (American Psychiatric Association, 2013). Dentre os termos comuns a todos os *embeddings* para ambas desordens, destacam-se “ideas” e “afraid”, cujo conjunto de termos associados por cada *embedding* apresentam comportamentos/emoções que podem ser típicos para a comorbidade, tais como “doubt” e “apprehensive” que representaria a apreensão e preocupação em evitar problemas (Ansiedade), bem como “avoid” e “frustrations”, que remeteria evitar dúvidas e frustrações (Depressão);

- *Diferenciadores*: além dos termos destacados pelos modelos especialistas, os diferenciadores agregaram termos diferentes à tomada de decisão. Por exemplo, o diferenciador CA-D destacou ainda os termos “anymore” e “effects” associados a “worry”, “afraid”, “reactions”, “relationships”, que podem ser associados a preocupações e medos significativos de estar em público, sentimentos comuns tanto na Ansiedade, quanto na Comorbidade.

d) Controle. Para usuários saudáveis, verificou-se a presença de termos presentes no DS Ansiedade e DS Depressão, porém em quantidades menores se comparado às amostras de usuários diagnosticados. Em relação às funções dos classificadores fracos, observou-se que:

- *Especialistas*: as amostras corretamente classificadas foram definidas pelos modelos especialistas, os quais votaram majoritariamente pela classe controle. Inspecionando essas amostras, verifica-se a presença de termos cujos significados estão associados a sentimentos típicos tanto da Ansiedade, quanto da Depressão. O termo “if”, exemplifica esse cenário, onde para alguns classificadores ele é destacado no DS Ansiedade, em outros no DS Depressão. Como já mencionado, no DS Ansiedade esse termo está associado à preocupação com situações futuras e dúvida. Já no DS Depressão ele está associado a “change”, “need” e “careful”, o que poderia indicar uma necessidade de mudança, mas difícil de praticar, situação típica do depressivo;

- *Diferenciadores*: como o esperado, os modelos diferenciadores divergiram entre as classes Ansiedade e Depressão. Nota-se também que, na maioria dos casos, os termos relevantes são similares aos identificados pelos classificadores especialistas, não sendo destacados novos termos.

6.5.3 Resultados: Amostras Classificadas com Erro

Nesta seção, examinamos dois tipos de erros: atribuição incorreta da condição de comorbidade a usuários diagnosticados, e confusão entre usuários saudáveis e com condições alvo. A Tabela 6.8 ilustra exemplos de *features* relevantes nas amostras classificadas incorretamente de acordo com cada classificador fraco, junto com os termos presentes em cada DS. Considerando que a atribuição incorreta envolvendo a condição de comorbidade é o erro mais comum em ambos os casos, e de particular interesse deste trabalho, focou-se especificamente nesta questão. Assim, as classes Depressão e Ansiedade da Tabela 6.8 representam o primeiro tipo de erro, e as classes Controle e Comorbidade, o segundo.

a) Usuários Diagnosticados: Atribuição incorreta da condição de Comorbidade.

- *Especialistas*: quando uma amostra de Ansiedade ou Depressão é classificada incorretamente, de modo geral verificou-se que os classificadores de ansiedade e depressão destacaram palavras similares em todos os DSs. Esse comportamento pode ser observado tanto nas amostras de Ansiedade, quanto de Depressão destacadas na Tabela 6.8. Observou-se que cada modelo classificou a instância segundo à condição alvo na qual é especialista, resultando no final uma mesma proporção de votos para Ansiedade e para Depressão.

  Com relação aos termos relevantes, diferente das amostras corretamente classificadas, verifica-se uma concentração maior de termos comuns aos DSs. Para as amostras de Ansiedade, por exemplo, tem-se a predominância do termo “worry” entre os termos comuns aos DSs. Esse termo está relacionado a “fear”, “avoid”, “panic” para o DS Ansiedade, e a “careful”, “concerns”, “cautious” no DS Depressão. Esse termo representa sintomas típicos tanto da Depressão quanto da Ansiedade. Já nas amostras de Depressão, é predominante o termo “feeling” que está associado a “fear”, “despair” no DS Ansiedade, e associado a “anxiety”, “depression”, “panic”

Figura 6.8: Lista de termos relevantes SHAP em amostras classificadas com erro

Condição Mental	Os 20 Termos SHAP mais Relevantes	Modelo	Termos Relevantes SHAP encontrados nos Dicionários de Sintomas (conforme cada classificador base)		
			Termos comuns aos Dicionários de Ansiedade e Depressão	Somente no Dicionário de Ansiedade	Somente no Dicionário de Depressão
Ansiedade	religious, my, because, make, sure, keep, head, count, low, enough, worry, why, should, school, stop, very, would, say, killing, off	L-CA2 C-CA3 H-CA1	make, my, should, sure, very, worry, would	possible, if, say, why, me	killing
		L-CD2 C-CD2 H-CD2	because, make, my, should, sure, very, worry, would	hope, if, say, why	killing
		C-CA-D1	how, keep, my, significant, sure, worry	make, very, hope	enough, if, because, should, killing
		C-CA-AD1	how, keep, my, significant, sure, worry	make, very, hope	enough, because, should, killing
		C-CD-AD1	because, make, my, should, sure, very, worry, would	if, rather, say, why, hope, how	killing
Depressão	how, rape, feeling, happened, know, stop, my, fault, made, said, could, take, because, instead, wish, change, use, have, me, yelling	L-CA2 C-CA3 H-CA1	because, feel, feeling, my, take, wish	going, know, could, instead, how	change
		L-CD2 C-CD2 H-CD2	because, feel, feeling, my, take, wish	could, going, how, instead, know	change
		C-CA-D1	fault, feeling, how, my	yelling, know, can	enough, because, could, change, instead
		C-CA-AD1	fault, feeling, how, my	yelling, know, can	enough, because, could, instead
		C-CD-AD1	because, feeling, my, take	yelling, know, could, me, use, how	change
Comorbidade	cloud, knowledge, can, trust, my, why, use, everything, forget, exact, something, would, take, computers, security, due, loss, problem, could, result	L-CA2 C-CA3 H-CA1	due, loss, my, result, something, take, would	could, why, everything, use	exact, problem
		L-CD2 C-CD2 H-CD2	due, loss, my, result, something, take, would	could, everything, use, why	problem, security, exact
		C-CA-D1	my	make, something, forget, can	if, security, exact, wrong
		C-CA-AD1	my	mak, something, forget, can	if, security, exact, wrong
		C-CD-AD1	due, loss, make, my, something, take, would	if, why, everything, use	exact
Controle	experiencing, opinion, getting, care, awesome, planned, because, really, will, often, go, scale, income, gods, lose, general, very, things, would, say	L-CA2 C-CA3 H-CA1	because, experiencing, my, things, too, very, would	really, will, say, able	getting
		L-CD2 C-CD2 H-CD2	because, experiencing, my, things, too, very, would	able, really, say, will, you	getting, income
		C-CA-D1	lose, my, myself	will, down, able, often, new	experiencing, because, income
		C-CA-AD1	lose, my, myself	down, able	enjoy, experiencing, because
		C-CD-AD1	because, experiencing, lose, my	really, will, down, able, me, harder, hard	much, enjoy, getting

no DS Depressão, sintomas comuns na condição de comorbidade;

- *Diferenciadores*: observou-se uma divergência entre esses modelos quanto à classificação das amostras. Novamente, notou-se que além dos termos já destacados como relevantes pelos modelos especialistas, esses classificadores destacaram novos termos. Por exemplo, o modelo CA-D classificou a amostra de Depressão corretamente. Além dos termos destacados pelos especialistas, nota-se os termos “enough"e “fault” destacados como relevantes por este modelo. Esses termos estão associados a “careful”, “failure”, “inability”, sentimentos típicos da depressão, o que poderia indicar um sentimento de receio em falhar por incapacidade de realizar tarefas;

- *Meta-learner*: verifica-se que mesmo nos casos em que a ansiedade foi identificada corretamente pelos modelos especialistas e os diferenciadores CA-D e CA-AD, a

amostra acabou recebendo ambos os rótulos (Depressão e Ansiedade). De modo similar, foram encontrados casos em que os modelos diferenciadores A-D e D-AD identificaram corretamente a classe Depressão, juntamente com os respectivos modelos especialistas, mas isso não foi suficiente para a correta classificação dessas amostras.

b) Erros de classificação envolvendo usuários saudáveis e diagnosticados.

Em relação aos termos relevantes relacionados aos classificadores fracos, nota-se no geral uma concentração maior de termos comuns entre os DSs Ansiedade e Depressão. Para a maioria das amostras inspecionadas, alguns desses termos relevantes poderiam ser relacionados com os sintomas da condição alvo, segundo o rótulo verdadeiro da classe, conforme mostra a Tabela 6.8. Mesmo assim, essas amostras não foram corretamente classificadas.

Ao inspecionar as amostras com erros, considerando as funções dos classificadores fracos, observou-se que:

1. Usuários saudáveis classificados como diagnosticado com Comorbidade:

   - *Especialistas*: todos votaram na condição alvo em que são especialistas;
   - *Diferenciadores*: divergiram quanto à classe da amostra, não havendo consenso;
   - *Meta-learner*: apenas refletiram as decisões dos classificadores fracos, segundo a maioria.

2. Usuários diagnosticados classificados como saudáveis:

   - *Especialistas*: uma unanimidade entre os classificadores especialistas a favor da classe Controle;
   - *Diferenciadores*: uma divergência desses modelos quanto à classificação da amostra entre Comorbidade e uma única condição alvo;
   - *Meta-learner*: mesmo nos casos em que os diferenciadores identificaram corretamente as amostras, ainda sim, a decisão final de classificação foi a favor da classe Controle.

6.5.4 Resultados: Recursos e seu Impacto na Variabilidade do Modelo *DAC Stacking*

Por fim, avaliou-se o impacto da variação arquitetural e de *embeddings* pré-treinados na variabilidade do modelo *DAC Stacking*. Para quase todas as amostras analisadas, verificou-se o mesmo comportamento em relação a cada variação de recurso. As Tabelas 6.13 e 6.14 ilustram um exemplo representativo considerando uma mesma amostra de Comorbidade corretamente classificada. Essas tabelas apresentam a relação das *features* influentes para a classificação em cada modelo especialista, de acordo com os valores SHAP.

Tabela 6.13: Variação Arquitetural (mesmo *embedding* pré-treinado)

Condição Mental	Os 20 Termos SHAP mais Relevantes	Modelo	Termos Relevantes SHAP encontrados nos Dicionários de Sintomas (conforme cada classificador base)		
			Termos comuns aos Dicionários de Ansiedade e Depressão	Somente no Dicionário de Ansiedade	Somente no Dicionário de Depressão
Comorbidade	my, mind, night, will, ideas, discussion, afraid, experience, cold, weird, consider, obviously, really trying, me, anymore, because, migraines, terrible, crying	L-CA2	ideas, mind, my, night,	afraid, really, having, will, cold,	experience, obviously, terrible,
		C-CA1	ideas, mind, my, night,	afraid, really, having, will, cold,	experience, obviously, terrible,
		H-CA1	ideas, mind, my, night,	afraid, really, having, will, cold,	experience, obviously, much
		L-CD2	ideas, mind, my, night,	afraid, having, will, cold, weird,	experience, much
		C-CD1	ideas, mind, my, night,	afraid, really, having, me, will,	experience, obviously, even,
		H-CD2	ideas, mind, my, night,	afraid, having, will, cold, weird	experience, obviously, much

Tabela 6.14: Variação de *Embeddings* Pré-treinados (somente Arquitetura CNN).

Condição Mental	Os 20 Termos SHAP mais Relevantes	Modelo	Termos Relevantes SHAP encontrados nos Dicionários de Sintomas (conforme cada classificador base)		
			Termos comuns aos Dicionários de Ansiedade e Depressão	Somente no Dicionário de Ansiedade	Somente no Dicionário de Depressão
Comorbidade	my, mind, night, will, ideas, discussion, afraid, experience, cold, weird, consider, obviously, really trying, me, anymore, because, migraines, terrible, crying	C-CA1	ideas, mind, my, night	afraid, really, having, will, cold	experience, obviously, terrible
		C-CA2	afraid, ideas, mind, my	can, will, cold, trying, weird	but, it, consider, obviously
		C-CA3			
		C-CD1	ideas, mind, my, night	afraid, really, having, me, will	experience, obviously, even
		C-CD2	afraid, crying, ideas, mind	can, will, cold, trying, weird	but, it, even, consider, obviously
		C-CD3	ideas	cold, weird, mind	

Quanto à variação de cada recurso, de modo geral, observou-se que:

- *Variação arquitetural*: comparando as diferentes arquiteturas usadas para formação dos classificadores especialistas, verificou-se que o conjunto de termos relevantes destacados é praticamente idêntico, como mostra a Tabela 6.13. Considerando os classificadores CA_i, nota-se que os modelos de arquitetura LSTM e CNN foram os que apresentaram maior similaridade em relação aos termos relevantes destacados nos dicionários de sintomas. Já nos classificadores CD_i, observa-se que a similaridade maior entre os termos relevantes ocorre entre as arquiteturas LSTM e Híbrida;

- *Variação de embeddings* pré-treinados: observou-se que tanto os classificadores especialistas em Ansiedade, quanto em Depressão apresentaram diferentes termos relevantes, conforme o *embedding* usado, como mostra a Tabela 6.14. Verifica-se que os modelos CA_i e CD_i apresentaram uma variabilidade maior entre os termos

relevantes quando formados usando os *embeddings* pré-treinados de propósito geral.

Assim, há evidências que o recurso que mais teve impacto na variabilidade do comitê foi a variação de *embeddings*.

6.5.5 Discussão

A avaliação qualitativa mostrou que as *features* destacadas pelos classificadores fracos são relevantes e oferecem *insights* sobre os transtornos alvo deste estudo. Dessa análise, obteve-se os seguintes *insights*:

- Usuários ansiosos estão mais relacionados ao termos representados no DS Ansiedade;
- Usuários depressivos estão mais relacionados aos termos do DS Depressão ou aos termos comuns a ambos os dicionários;
- Usuários diagnosticados com a condição de comorbidade estão mais relacionados aos termos comuns dos DSs;
- Usuários saudáveis estão eventualmente relacionados a termos de todos os dicionários, mas em quantidades menores que as observadas nos usuários diagnosticados.

A relação das *features* influentes com os DSs permitiu constatar que as características dos Transtornos de Ansiedade estão presentes em algum nível de intensidade em todos os usuários. Como mostrou a análise de amostras classificadas incorretamente, essa característica contribuiu para os erros de classificação observados em relação aos usuários diagnosticados, onde todos os modelos *DAC Stacking* apresentam baixa precisão para identificação da Comorbidade.

A análise das amostras classificadas com erro também permitiu identificar que os classificadores diferenciadores, mesmo apresentando desempenho inferior aos modelos especialistas, são capazes de contribuir para a variabilidade do modelo *DAC Stacking*. Nessa análise, constatou-se que os modelos diferenciadores apresentam *features* influentes relevantes para identificação da classe alvo e diferentes das identificadas pelos classificadores especialistas.

A análise dos erros permitiu constatar também que a consolidação realizada pelo modelo *meta-learner* deve ser melhorada, uma vez que para todas as classes foram encontrados exemplos de erros nos quais a maioria dos classificadores fracos identificaram corretamente a amostra, mas mesmo assim ela foi classificada errada.

Finalmente, a análise comparativa para avaliação do impacto de variar os recursos (arquitetura e/ou *embeddings* pré-treinados) mostrou que o maior impacto de variabilidade é resultante da variação de *embeddings* e não das arquiteturas. Ainda, um conjunto maior de *features* relevantes foram apresentados pelos modelos que usavam os *embedding* pré-treinados de propósito geral.

6.6 Experimento #5: BERT e a Classificação Multi-tarefa

Este experimento tem por objetivo comparar a performance do *DAC* Stacking com modelos BERT, o atual estado da arte para solução de problemas de classificação envolvendo conteúdo textual. Este experimento tem por objetivo responder à QP6, i.e. *"Como a solução DAC Stacking proposta se compara a soluções estado da arte no tratamento de problemas na área de Processamento de Linguagem Natural (PLN)?"*. Assim, desenvolvemos modelos BERT para o mesmo problema de classificação multirrótulo, e comparamos seu desempenho com o modelo *DAC Stacking DT LCH-C*, que apresentou os melhores resultados.

6.6.1 Método

Para estabelecer o comparativo, desenvolvemos 4 modelos baseados em BERT, usando as duas estratégias possíveis:

a) Modelo Principal (MP): um modelo denominado *BERT MP* foi gerado pelo ajuste fino de BERT, configurado para usar a mesma camada de saída empregada nos modelos multirrótulo desse trabalho (Seção 5.2);

b) *Embedding*: desenvolvemos modelos denominado *BERT L*, *BERT C* e *BERT H* usando BERT como uma camada de *embeddings* das configurações de melhor desempenho das arquiteturas multirrótulo LSTM, CNN e Híbrida, respectivamente.

Todos estes modelos foram treinados e validados com o conjunto de dados SMHD

A-D-AD, utilizando o BERT Base Uncased. Como a geração de modelos baseados em BERT demanda um poder computacional muito grande para seu treinamento e execução, utilizamos o serviço Google Colab[1], o qual permite alocar servidores com poder computacional dedicado. Os modelos foram codificados usando Jupyter Notebook e estão disponíveis junto dos respectivos código fonte no repositório público.

Os modelos BERT foram analisados quantitativamente usando as mesmas métricas utilizadas na avaliação dos modelos *DAC Stacking* no Experimento 3 (Seção 6.4). Os resultados reportados correspondem à performance média considerando 10 repetições para cada teste. Para verificação de significância estatística foi usado o Teste Student T bicaudal pareado ($alpha = 0.5$). Como hipótese nula ($p - value >= 0.05$), formulou-se que não há diferença significativa de performance entre duas arquiteturas comparadas. Essa análise foi realizada entre o modelo BERT MP e as variações BERT propostas, bem como entre o modelo BERT de melhor performance e o modelo *DAC Stacking DT LCH-C*.

Finalmente, realizou-se uma análise de erros do modelo BERT de melhor desempenho, comparando estes resultados ao modelo *DAC Stacking DT LCH-C*.

6.6.2 Resultados

A Tabela 6.15 apresenta o desempenho médio dos modelos BERT desenvolvidos, junto ao do *DAC Stacking DT LCH-C*. Os resultados dos testes estatísticos são detalhados no Apêndice F, juntamente com a relação de valores de média e desvio padrão para cada experimento. Em relação aos modelos BERT, verifica-se que os modelos *BERT L* e *BERT H* apresentaram os melhores resultados, não havendo diferença estatisticamente significativa em nenhuma das métricas analisadas. Já o modelo *BERT C* foi o que apresentou o pior desempenho em todas as métricas. Considerando a medida EMR, por exemplo, *BERT C* apresentou performance inferior entre 39 e 42 pp quando comparado aos demais modelos.

Tabela 6.15: Modelos BERT: Performance Média

Modelo	Multirrótulo		Control			Ansiedade			Depressão			Comorbidade		
	EMR	HL	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT MP	0,42	0,28	0,72	0,73	0,72	0,64	0,61	0,63	0,62	0,50	0,54	0,34	0,54	0,41
BERT L	0,45	0,28	0,73	0,70	0,72	0,62	0,72	0,66	0,57	0,76	0,65	0,32	0,75	0,45
BERT C	0,03	0,39	0,83	0,05	0,10	0,78	0,04	0,07	0,64	0,04	0,07	0,37	0,04	0,07
BERT H	0,44	0,28	0,75	0,67	0,71	0,61	0,74	0,67	0,56	0,77	0,65	0,32	0,77	0,45
DAC Stacking DT	0,42	0,28	0,74	0,63	0,68	0,62	0,73	0,67	0,58	0,72	0,64	0,33	0,75	0,46

Comparando-se o modelo *BERT MP* e suas variações *BERT L* e *BERT H*, observa-

[1]Google Colab: https://colab.research.google.com/notebooks/intro.ipynb

se uma melhor performance dos modelos que usam BERT como *embeddings* em quase todas as métricas. Os resultados são superiores (2 a 3 pp) para a métrica EMR, e comparáveis para a métrica HL. Em termos de medida F, verifica-se que as variações *BERT L* e *BERT H* são significativamente superiores ao modelo *BERT MP* para as classes Ansiedade (3 a 4 pp) e Depressão (11 pp). Embora apresentem performance superior em medida F para as classes Comorbidade e Controle, essas diferenças não foram significativas.

Considerando que os modelos *BERT L* e *BERT H* são estatisticamente similares, a análise de erros foi realizada considerando o modelo *BERT H*. O erro de predizer um usuário saudável como diagnosticado foi observado em 11% das amostras de teste (4% para Depressão e 7% para Comorbidade). Já o erro inverso (predizer usuário diagnosticado como saudável) foi observado em 9% das amostras, distribuída em 3% para Ansiedade, 4% para Depressão e 2% para Comorbidade.

Quanto ao erro para diferenciar entre os usuários diagnosticados, novamente o erro mais frequente é predizer um usuário com um único transtorno como apresentando comorbidade. Esse erro ocorre para 14% dos usuários ansiosos e 12% dos usuários depressivos. Também foram registrados erros nas seguintes condições: predizer usuário ansioso como depressivo (2%) e predizer usuário com comorbidade como depressivo (3%). Não foram observados casos envolvendo usuários depressivos classificados incorretamente como ansiosos.

Em relação aos modelos BERT, *DAC Stacking DT LCH-C* apresentou um resultado levemente inferior para a métrica EMR, e superior para muitas métricas relacionadas às condições alvo. No entanto, apenas a medida F para a classe Comorbidade apresentou diferença estatística significativa de 1 pp, se comparado aos modelos *BERT L* e *BERT H*.

Considerando a análise de erros, algumas diferenças foram observadas entre os modelos *BERT H* e *DAC Stacking DT LCH-C*. Em relação ao erro de diferenciar entre usuários saudáveis e diagnosticados, verifica-se que o modelo *DAC Stacking DT LCH-C* tem uma taxa de 3 pp menor de erros, e 1 pp para o erro inverso.

Já para o erro mais frequente referente à condição de comorbidade, o modelo *DAC Stacking DT LCH-C* apresentou um desempenho sutilmente melhor. Comparativamente, a taxa de predizer usuários depressivos como com condição de comorbidade é 2 pp menor, e a de usuários ansiosos, 1 pp menor. Já o modelo *BERT H* não apresentou o erro de predizer usuário depressivo como ansioso.

6.6.3 Discussão

Em relação ao uso de BERT, os resultados desse experimento permitem concluir que, para a tarefa de classificação multirrótulo de usuários com as condições alvo deste estudo, seu uso como camada de *embeddings* foi o que apresentou melhor desempenho. Entre as variações experimentadas, os modelos que combinam o *embedding* BERT às arquiteturas LSTM e Híbrida exploradas neste trabalho apresentaram melhores resultados.

O comparativo de performance entre os classificadores *DAC Stacking DT LCH-C* e *BERT H* mostrou que esses modelos são estatisticamente similares considerando o conjunto de métricas avaliado. No entanto, a análise de erros mostrou que a solução proposta nesse trabalho apresentou desempenho sutilmente melhor, reduzindo o erro para identificar entre usuários saudáveis e diagnosticados.

Os resultados desse experimento validam tanto a contribuição da solução de comitê representada por *DAC Stacking*, como as contribuições de um modelo de representação de linguagem estado da arte como BERT. Em particular, apontam direções futuras para este trabalho no tocante ao uso de *embeddings* BERT como uma alternativa para aumentar a performance dos modelos diferenciadores. Além disso, como visto no Experimento #4 (Seção 6.5.5), a inclusão de diferentes *embeddings* pré-treinados para formação dos classificadores fracos é uma alternativa eficaz para promover ganho de variabilidade no comitê.

7 CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho foi proposto o modelo *DAC Stacking* para determinar a classificação de usuários diagnosticados com ansiedade, depressão e a condição de comorbidade entre esses transtornos, a partir de dados textuais da rede social Reddit. A formação do comitê foi baseada em um extenso processo de experimentação, o qual avaliou a função dos classificadores fracos que compõem o comitê, arquiteturas de aprendizado profundo e *embeddings*. A avaliação foi feita em termos quantitativos e qualitativos. Em relação à avaliação qualitativa, propusemos relacionar *features* influentes na classificação segundo a métrica valores SHAP com sintomas do manual DSM-5.

A solução de comitê proposta mostrou-se promissora para a identificação dos transtornos alvo deste estudo. Os resultados mostraram que o modelo *DAC Stacking* proposto é mais efetivo para a solução deste problema que o uso de um único modelo de aprendizado profundo multirrótulo. A realização de um extenso conjunto de experimentos permitiu compreender as fortalezas e limites da solução proposta.

De uma forma geral, explorou-se de forma bastante ampla a influência dos diferentes hiperparâmetros que permitiriam a melhoria de desempenho de cada classificador fraco. Dentre de todas variações propostas para os modelos *DAC Stacking*, nenhuma se destacou em termos de desempenho. Mas a análise qualitativa permitiu compreender a contribuição de cada recurso na variabilidade do comitê e as razões de alguns comportamentos.

Dentre todas as variações propostas para a composição do *DAC Stacking*, o uso de diferentes *embeddings* pré-treinados de propósito geral foi o que resultou no maior ganho de variabilidade. A variação de função para os classificadores fracos do *Nível 0*, divididos em especialistas e diferenciadores de condições alvo, também mostrou-se efetiva para recuperação de *features* influentes diferentes. Contrariando a expectativa inicial, o uso de arquiteturas com diferentes premissas de aprendizado (LSTM, CNN e modelos híbridos), embora apresente bons resultados para classificação individual de cada condição alvo, pouco contribuíram para a variabilidade do modelo de comitê.

A análise qualitativa permitiu constatar que as *features* influentes para a classificação são representativas dos sintomas relacionados a cada condição alvo. Ainda, destacou que os traços de ansiedade estão presente em todos os usuários (diagnosticados e saudáveis), o que contribui à dificuldade apresentada pelos modelos *DAC Stacking* para distinguir entre os usuários diagnosticados. Além dos *insights* revelados, essa análise contribuiu

para compreensão das limitações do modelo *DAC Stacking*, em particular relacionada à precisão da classificação da condição de comorbidade.

A principal limitação do *DAC Stacking* é a classificação errônea de usuários envolvendo a comorbidade. Contudo, é importante destacar que todas as soluções às quais *DAC Stacking* foi comparada apresentaram resultado inferior em relação a este problema. A análise de erros revelou que a baixa precisão da classificação da comorbidade está relacionada à situação em que usuários diagnosticados com uma única condição são identificados com comorbidade. As análises mostraram duas razões principais para este tipo de erro. Por um lado, notou-se que apesar de diferentes premissas de aprendizado das arquiteturas LSTM, CNN e Híbrida usadas para compor o *Nível 0* dos modelos *DAC Stacking*, elas resultam em conjuntos similares de *features* relevantes para a tomada de decisão quanto à classificação das amostras, indicando que os padrões aprendidos por elas são similares. Por outro lado, as avaliações mostraram que o modelo *meta-learner* deve ser melhorado para consolidação das previsões corretamente realizadas pelos classificadores fracos.

Embora os classificadores diferenciadores apresentem variabilidade em relação aos termos relevantes destacados para a classificação, sua presença contribuiu pouco para o ganho performance do modelo *DAC Stacking*, devido à baixa performance alcançada por esses modelos. A etapa de experimentação para formação dos diferenciadores esgotou as arquiteturas testadas quanto à parametrização por ajuste fino e manual, mas mesmo assim o desempenho final desses modelos foi muito inferior aos especialistas. Disso, constata-se que a diferenciação entre as condições alvo desse estudo apresenta padrões sutis que as arquiteturas exploradas não foram capazes de aprender.

Os experimentos realizados com BERT mostraram que o mero uso desta solução estado da arte para problemas de NLP como um modelo resultante do ajuste fino não produz uma solução superior a *DAC Stacking*. Também permitiram vislumbrar alternativas para limitações de *DAC Stacking* através de novos experimentos com os diferenciadores, e o uso de BERT como *embeddings*.

Resultados preliminares desta pesquisa foram publicados no Simpósio Brasileiro de Banco de Dados (SOUZA; NOBRE; BECKER, 2020), tendo recebido a distinção de melhor artigo deste evento (Prêmio José Mauro Volkmer Castilho).

Em relação aos trabalhos futuros, destacam-se entre outros: a) desenvolvimento de classificadores do tipo diferenciador com melhor desempenho; b) exploração de alternativas para o *meta-learner*, possivelmente através de um comitê do tipo *cascading*;

c) investigação de como incorporar a técnica BERT (e suas variações) na solução *DAC Stacking*; d) uso de mecanismos de atenção para a consolidação dos resultados do comitê; e) extensão desta proposta a outras redes sociais ou a situações específicas (tais como o contexto da pandemia), entre outros.

REFERÊNCIAS

ALTHOFF, T.; CLARK, K.; LESKOVEC, J. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. **Transactions of the Association for Computational Linguistics**, v. 4, p. 463–476, 2016. Available from Internet: <https://www.aclweb.org/anthology/Q16-1033>.

American Psychiatric Association. **Diagnostic and statistical manual of mental disorders: DSM-5**. 5th ed.. ed. Washington, DC: Autor, 2013. 155,189-190 p.

ARAUJO, C.; NETO, F. L. A Nova Classificação Americana Para os Transtornos Mentais: o DSM-5. **Revista Brasileira de Terapia Comportamental e Cognitiva**, scielopepsic, v. 16, p. 67 – 82, 04 2014. ISSN 1517-5545. Available from Internet: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1517-55452014000100007&nrm=iso>.

BAGROY, S.; KUMARAGURU, P.; CHOUDHURY, M. D. A social media based index of mental well-being in college campuses. In: **Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2017. (CHI '17), p. 1634–1646. ISBN 9781450346559. Available from Internet: <https://doi.org/10.1145/3025453.3025909>.

BENTON, A.; MITCHELL, M.; HOVY, D. Multitask learning for mental health conditions with limited social media data. In: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**. Valencia, Spain: Association for Computational Linguistics, 2017. p. 152–162. Available from Internet: <https://www.aclweb.org/anthology/E17-1015>.

BRADLEY, M. M. et al. **Affective Norms for English Words (ANEW): Instruction manual and affective ratings**. 1999.

BRITZ, D. **Implementing a CNN for Text Classification in TensorFlow**. 2016. Available from Internet: <http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/>.

CACHEDA, F. et al. Early detection of depression: Social network analysis and random forest techniques. **J Med Internet Res**, v. 21, n. 6, p. e12554, Jun 2019. ISSN 1438-8871. Available from Internet: <http://www.jmir.org/2019/6/e12554/>.

CHO, K. et al. On the properties of neural machine translation: Encoder–decoder approaches. In: **Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 103–111. Available from Internet: <https://www.aclweb.org/anthology/W14-4012>.

CHOLLET, F. What is deep learning? In: **Deep Learning with Python**. [S.l.]: Manning, 2017. chp. 1, p. 8–22;94–96,102–104,123;184–185,196–197,202–206,215–216;264–266. ISBN 9781617294433.

CHOUDHURY, M. D.; COUNTS, S.; HORVITZ, E. Predicting postpartum changes in emotion and behavior via social media. In: **Proceedings of the SIGCHI Conference**

on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2013. (CHI '13), p. 3267–3276. ISBN 9781450318990. Available from Internet: <https://doi.org/10.1145/2470654.2466447>.

CHOUDHURY, M. D. et al. Characterizing and predicting postpartum depression from shared facebook data. In: **Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work amp; Social Computing**. New York, NY, USA: Association for Computing Machinery, 2014. (CSCW '14), p. 626–638. ISBN 9781450325400. Available from Internet: <https://doi.org/10.1145/2531602.2531675>.

CHOUDHURY, M. D. et al. Social media participation in an activist movement for racial equality. In: **ICWSM**. [S.l.: s.n.], 2016. p. 92–101.

CHOUDHURY, M. D. et al. Discovering shifts to suicidal ideation from mental health content in social media. In: **Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2016. (CHI '16), p. 2098–2110. ISBN 9781450333627. Available from Internet: <https://doi.org/10.1145/2858036.2858207>.

CHOUDHURY, M. D. et al. Gender and cross-cultural differences in social media disclosures of mental illness. In: **Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing**. New York, NY, USA: Association for Computing Machinery, 2017. (CSCW '17), p. 353–369. ISBN 9781450343350. Available from Internet: <https://doi.org/10.1145/2998181.2998220>.

CICCHETTI, D. V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. **Psychological Assessment**, v. 6, p. 284–290, 1994.

COHAN, A. et al. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In: BENDER, E. M.; DERCZYNSKI, L.; ISABELLE, P. (Ed.). **Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018**. Association for Computational Linguistics, 2018. p. 1485–1497. Available from Internet: <https://www.aclweb.org/anthology/C18-1126/>.

COPPERSMITH, G.; DREDZE, M.; HARMAN, C. Quantifying mental health signals in Twitter. In: **Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality**. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. p. 51–60. Available from Internet: <https://www.aclweb.org/anthology/W14-3207>.

COPPERSMITH, G. et al. From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses. In: **Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 1–10. Available from Internet: <https://www.aclweb.org/anthology/W15-1201>.

COPPERSMITH, G.; HARMAN, C.; DREDZE, M. Measuring post traumatic stress disorder in twitter. In: **ICWSM**. The AAAI Press, 2014. (ICWSM'14, v. 2), p. 23–45. Available from Internet: <http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.html#HuttoG14>.

DENG, L.; LIU, Y. **Deep Learning in Natural Language Processing**. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2018. ISBN 9789811052088.

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, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Available from Internet: <https://www.aclweb.org/anthology/N19-1423>.

DUTTA, S.; MA, J.; CHOUDHURY, M. D. Measuring the impact of anxiety on online social interactions. In: **ICWSM**. [S.l.]: The AAAI Press, 2018. (ICWSM'18).

FELLBAUM, C. (Ed.). **WordNet: an electronic lexical database**. [S.l.]: MIT Press, 1998.

GAMA, J. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Grupo Gen - LTC, 2011. ISBN 9788521618805. Available from Internet: <https://books.google.com.br/books?id=4DwelAEACAAJ>.

GANEGEDARA, T. **Natural Language Processing with TensorFlow: Teach language to machines using Python's deep learning library**. Packt Publishing, 2018. ISBN 9781788477758. Available from Internet: <https://books.google.com.br/books?id=LHxeDwAAQBAJ>.

GIUNTINI, F. T. et al. A review on recognizing depression in social networks: challenges and opportunities. **Journal of Ambient Intelligence and Humanized Computing**, JMIR Publications, n. 1868-5145, 2020. Available from Internet: <https://doi.org/10.1007/s12652-020-01726-4>.

GKOTSIS, G. et al. Characterisation of mental health conditions in social media using informed deep learning. In: . A Nature Research Journal, 2017. v. 7, p. 2045–2322. Available from Internet: <https://doi.org/10.1038/srep45141>.

GOLDBERG, Y. **Neural Network Methods for Natural Language Processing**. San Rafael, CA: Morgan & Claypool, 2017. (Synthesis Lectures on Human Language Technologies, v. 37). ISSN 1947-4040. ISBN 978-1-62705-298-6.

GONZALES, A. L.; HANCOCK, J. T.; PENNEBAKER, J. W. Language style matching as a predictor of social dynamics in small groups. Communication Research, v. 37, n. 1, p. 3–19, 2010.

GRUDA, D.; HASAN, S. Feeling anxious? perceiving anxiety in tweets using machine learning. **Computers in Human Behavior**, v. 98, p. 245 – 255, 2019. ISSN 0747-5632. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0747563219301608>.

HALFIN, A. Depression: the benefits of early and appropriate treatment. **The American journal of managed care**, v. 13, n. 4 Suppl, p. S92–7, 2007.

HAMILTON, M. Development of a rating scale for primary depressive illness. **British Journal of Social and Clinical Psychology**, v. 6, n. 4, p. 278–296, 1967. Available from

Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8260.1967.tb00530.x>.

HART, S. Shapley Values. In: Eatwell, J. and Milgate, M. and Newman, P. (Ed.). **Game Theory**. The New Palgrave: Palgrave Macmillan, London, 1989. p. 210–216. ISBN 978-0-333-49537-7.

He, K. et al. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 770–778.

HEALTH, N. I. for; EXCELLENCE, C. **Common mental health problems: identification and pathways to care**. 2011. Available from Internet: <https://www.nice.org.uk/guidance/cg123/resources/common-mental-health-problems-identification-and-pathways-to-care-pdf-35109448223173>.

HIRSCHFELD, R. The comorbidity of major depression and anxiety disorders: Recognition and management in primary care. **Prim Care Companion J Clin Psychiatry**, v. 3, n. 244-254, 12 2001.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Comput.**, MIT Press, Cambridge, MA, USA, v. 9, n. 8, p. 1735–1780, nov. 1997. ISSN 0899-7667. Available from Internet: <https://doi.org/10.1162/neco.1997.9.8.1735>.

Hu, Q. et al. Predicting depression of social media user on different observation windows. In: **2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)**. [S.l.: s.n.], 2015. v. 1, p. 361–364.

HUTTO, C. J.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: ADAR, E. et al. (Ed.). **ICWSM**. The AAAI Press, 2014. ISBN 978-1-57735-659-2. Available from Internet: <http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.html#HuttoG14>.

IRELAND, M.; ISERMAN, M. Within and between-person differences in language used across anxiety support and neutral Reddit communities. In: **Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic**. New Orleans, LA: Association for Computational Linguistics, 2018. p. 182–193. Available from Internet: <https://www.aclweb.org/anthology/W18-0620>.

ISLAM, M. R. et al. Depression detection from social network data using machine learning techniques. In: . [s.n.], 2018. v. 6, p. 8. Available from Internet: <https://doi.org/10.1007/s13755-018-0046-0>.

IVE, J. et al. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In: **Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic**. New Orleans, LA: Association for Computational Linguistics, 2018. p. 69–77. Available from Internet: <https://www.aclweb.org/anthology/W18-0607>.

Keumhee Kang; Chanhee Yoon; Eun Yi Kim. Identifying depressive users in twitter using multimodal analysis. In: **2016 International Conference on Big Data and Smart Computing (BigComp)**. [S.l.: s.n.], 2016. p. 231–238.

LI, L. et al. A system for massively parallel hyperparameter tuning. In: DHILLON, I.; PAPAILIOPOULOS, D.; SZE, V. (Ed.). **Proceedings of Machine Learning and Systems**. [s.n.], 2020. v. 2, p. 230–246. Available from Internet: <https://proceedings.mlsys.org/paper/2020/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf>.

LIN, H. et al. User-level psychological stress detection from social media using deep neural network. In: **Proceedings of the 22nd ACM International Conference on Multimedia**. New York, NY, USA: Association for Computing Machinery, 2014. (MM '14), p. 507–516. ISBN 9781450330633. Available from Internet: <https://doi.org/10.1145/2647868.2654945>.

LOVEYS, K. et al. Small but mighty: Affective micropatterns for quantifying mental health from social media language. In: **Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality**. Vancouver, BC: Association for Computational Linguistics, 2017. p. 85–95. Available from Internet: <https://www.aclweb.org/anthology/W17-3110>.

LUNDBERG, S. M.; LEE, S. A unified approach to interpreting model predictions. In: GUYON, I.; LUXBURG, U. von; ALLI et (Ed.). **Advances in Neural Information Processing Systems : Proc. of the 30th Annual Conf. on Neural Information Processing Systems (NIPS)**. [S.l.: s.n.], 2017. p. 4765–4774.

MANIKONDA, L.; CHOUDHURY, M. D. Modeling and understanding visual attributes of mental health disclosures in social media. In: **Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2017. (CHI '17), p. 170–181. ISBN 9781450346559. Available from Internet: <https://doi.org/10.1145/3025453.3025932>.

MANN, P.; PAES, A.; MATSUSHIMA, E. H. See and read: Detecting depression symptoms in higher education students using multimodal social media data. In: **ICWSM**. [s.n.], 2020. v. 14, n. 1, p. 440–451. Available from Internet: <https://aaai.org/ojs/index.php/ICWSM/article/view/7313>.

MATERO, M. et al. Suicide risk assessment with multi-level dual-context language and BERT. In: **Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 39–44. Available from Internet: <https://www.aclweb.org/anthology/W19-3005>.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: BENGIO, Y.; LECUN, Y. (Ed.). **1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings**. [s.n.], 2013. Available from Internet: <http://arxiv.org/abs/1301.3781>.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. [S.l.]: The MIT Press, 2012. ISBN 0262018020, 9780262018029.

PARK, M.; MCDONALD, D. W.; CHA, M. Perception differences between the depressed and non-depressed users in twitter. In: KICIMAN, E. et al. (Ed.). **ICWSM**. [S.l.]: The AAAI Press, 2013. ISBN 978-1-57735-610-3.

PARK, S. et al. Manifestation of depression and loneliness on social networks: A case study of young adults on facebook. In: **Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing**. New York, NY, USA: Association for Computing Machinery, 2015. (CSCW '15), p. 557–570. ISBN 9781450329224. Available from Internet: <https://doi.org/10.1145/2675133.2675139>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Available from Internet: <https://www.aclweb.org/anthology/D14-1162>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Empirical Methods in Natural Language Processing (EMNLP)**. [s.n.], 2014. p. 1532–1543. Available from Internet: <http://www.aclweb.org/anthology/D14-1162>.

PETERS, M. et al. Deep contextualized word representations. In: **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. Available from Internet: <https://www.aclweb.org/anthology/N18-1202>.

PETERS, M. E. et al. Deep contextualized word representations. In: **Proc. of NAACL**. [S.l.: s.n.], 2018.

PREOŢIUC-PIETRO, D. et al. The role of personality, age, and gender in tweeting about mental illness. In: **Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 21–30. Available from Internet: <https://www.aclweb.org/anthology/W15-1203>.

PRIMACK, B. et al. Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among u.s. young adults. **Computers in Human Behavior**, v. 69, p. 1–9, 04 2017.

RADLOFF, L. S. The ces-d scale: A self-report depression scale for research in the general population. **Applied Psychological Measurement**, v. 1, n. 3, p. 385–401, 1977. Available from Internet: <https://doi.org/10.1177/014662167700100306>.

SECAD, R. **DSM-5: indispensável para diagnóstico de transtornos mentais**. SECAD, Artmed, 2018. Available from Internet: <https://www.secad.com.br/blog/psiquiatria/dsm-5-diagnostico-transtornos-mentais/>.

SHARMA, E.; CHOUDHURY, M. D. Mental health support and its relationship to linguistic accommodation in online communities. In: **Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2018. (CHI '18), p. 1–13. ISBN 9781450356206. Available from Internet: <https://doi.org/10.1145/3173574.3174215>.

SHEN, J. H.; RUDZICZ, F. Detecting anxiety through Reddit. In: **Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality**. Vancouver, BC: Association for Computational Linguistics, 2017. p. 58–65. Available from Internet: <https://www.aclweb.org/anthology/W17-3107>.

SOUZA, V. B.; NOBRE, J. C.; BECKER, K. Characterization of Anxiety, Depression, and their Comorbidity from Texts of Social Networks. In: **Anais do XXXV Simpósio Brasileiro de Banco de Dados**. Porto Alegre, RS, Brasil: SBC, 2020. Available from Internet: <https://sbbd.org.br/2020/wp-content/uploads/sites/13/2020/09/Characterizing-Anxiety-ST7.pdf>.

SPIEGEL, M. R. et al. **Estatistica**. 4. ed. [S.l.]: Bookman, 2009.

TADESSE, M. M. et al. Detection of depression-related posts in reddit social media forum. **IEEE Access**, v. 7, p. 44883–44893, 2019.

TADESSE, M. M. et al. Detection of depression-related posts in reddit social media forum. **IEEE Access**, v. 7, p. 44883–44893, 2019.

TILLER, J. Depression and anxiety. **The Medical journal of Australia**, v. 199, p. S28–31, 09 2013.

TSUGAWA, S. et al. Recognizing depression from twitter activity. In: **Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2015. (CHI '15), p. 3187–3196. ISBN 9781450331456. Available from Internet: <https://doi.org/10.1145/2702123.2702280>.

TUCHLINKSKI, C. **Depressão será a doença mental mais incapacitante do mundo até 2020**. 2018. Available from Internet: <https://emais.estadao.com.br/noticias/bem-estar,depressao-sera-a-doenca-mental-mais-incapacitantes-do-mundo-ate-2020,70002542030>.

VASWANI, A. et al. Attention is all you need. In: **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

WHO, W. H. O. **Depression and Other Common Mental Disorders: Global Health Estimates**. Geneva: Word Health Organization, 2017. Available from Internet: <https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf>.

WONGKOBLAP, A.; VADILLO, M.; CURCIN, V. Researching mental health disorders in the era of social media: Systematic review. **Journal of Medical Internet Research**, JMIR Publications, v. 19, n. 6, 2017. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5509952/>.

YATES, A.; COHAN, A.; GOHARIAN, N. Depression and self-harm risk assessment in online forums. In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 2968–2978. Available from Internet: <https://www.aclweb.org/anthology/D17-1322>.

YATES, A.; GOHARIAN, N. Adrtrace: Detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In: SERDYUKOV, P. et al. (Ed.). **Advances in Information Retrieval**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 816–819. ISBN 978-3-642-36973-5.

YOM-TOV, E.; GABRILOVICH, E. Postmarket drug surveillance without trial costs: Discovery of adverse drug reactions through large-scale analysis of web search queries. **J Med Internet Res**, v. 15, n. 6, p. e124, Jun 2013. ISSN 14388871. Available from Internet: <http://www.jmir.org/2013/6/e124/>.

ZHANG, C.; MA, Y. **Ensemble Machine Learning: Methods and Applications**. [S.l.]: Springer Publishing Company, Incorporated, 2012. ISBN 1441993258.

APÊNDICE A — VARIAÇÕES EXPLORADAS PARA A CAMADA DE *EMBEDDINGS*

A etapa inicial de experimentação para definir a camada de *embeddings* para os classificadores especialistas usando arquitetura LSTM explorou os formatos (1) ausência da camada de *embeddings* e o (2) uso de *embeddings* randômicos, no qual os pesos para os *embeddings* são iniciados aleatoriamente e ajustados durante o treinamento do modelo. A Tabela A.1 apresenta a configuração e resultados obtidos nessa etapa de experimentos. De modo similar, a Tabela A.2 apresenta a configuração e resultados obtidos para a experimentação inicial envolvendo a formação dos classificadores especialistas usando a arquitetura CNN, quanto ao uso de *embeddings* randômicos vs. pré-treinados de propósito geral (Glove 6B).

Tabela A.1: Uso de *Embeddings*: Exploração Inicial para arquitetura LSTM

Condição Mental	**Parametrização Rede Neural**	***Embedding***	**P**	**R**	**F1**
Ansiedade	3 camadas LSTM Return sequence = True Dropout de 20% Aprendizado Não Estático	Glove 6B	0,60	0,64	0,57
		Randômico	0,55	0,54	0,55
Depressão		Glove 6B	0,66	0,65	0,67
		Randômico	0,57	0,55	0,58
Comorbidade		Glove 6B	0,55	0,59	0,58
		Randômico	0,51	0,61	0,51

Tabela A.2: Uso de *Embeddings*: Exploração Inicial para arquitetura CNN

Condição Mental	**Parametrização Rede Neural**	***Embedding***	**P**	**R**	**F1**
Ansiedade	Kernel size = 5, Dropout = 50% Filters = 250	Glove 6B	0,80	0,60	0,68
		Randômico	0,59	0,68	0,63
Depressão	Kernel size = 4, Dropout = 20% Filters = 250	Glove 6B	0,81	0,65	0,72
		Randômico	0,58	0,70	0,64
Comorbidade	Kernel size = 4, Dropout = 50% Filters = 100	Glove 6B	0,81	0,52	0,63
		Randômico	0,55	0,54	0,54

APÊNDICE B — EXPERIMENTOS COM A FUNÇÃO DE INICIALIZAÇÃO DO *KERNEL* EM REDES NEURAIS

Esse experimento explora a variação da função de inicialização do *kernel* da rede, segundo as opções *Glorot Uniform* e *LeCun Uniform*. Para tanto, o experimento foi estruturado nas seguintes etapas:

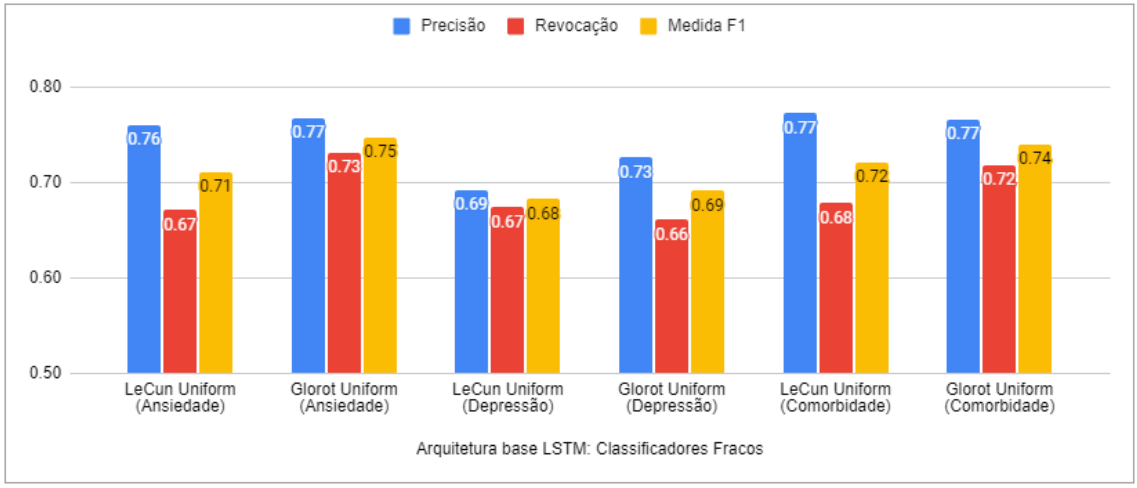
1. Seleção dos classificadores fracos: foram considerados os modelos especialistas em cada condição alvo, gerados tanto com a arquitetura LSTM, quanto CNN. A definição da topologia adotada para cada arquitetura é detalhada nas Seções 5.5.1.1 (Tabela 5.2) e 5.5.1.2 (Tabela 5.3), respectivamente.

2. Execução do conjunto de testes: para cada classificador fraco, dois experimentos foram definidos, conforme o tipo de função usada para inicialização do *kernel* da rede. Ao todo, doze experimentos diferentes foram realizados. Para cada experimento, dez repetições foram executadas com a finalidade de computar a performance de classificação para os modelos gerados, considerando seu desempenho para o conjunto de dados de teste.

3. Análise de Performance: foi empregado o teste estatístico *T Student* (Bicaudal pareado) para comparar a performance dos modelos quanto às métricas precisão, revocação e medida F. Desse modo, cada análise comparou o desempenho entre modelos de mesma arquitetura e função de classificação, porém usando funções de inicialização do *kernel* diferentes.

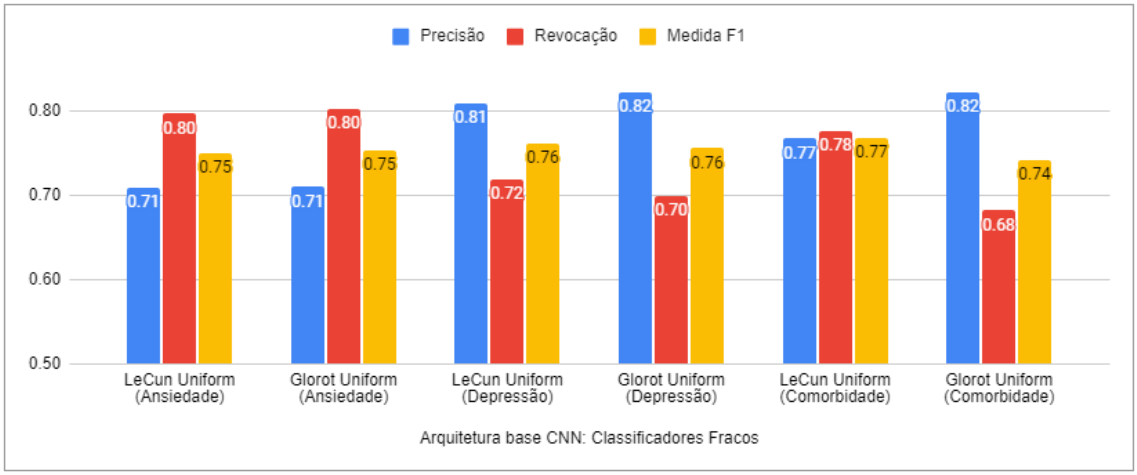
A Figura B.1 compara a média para os resultados produzidos pelas funções Glorot e LeCun. Considerando os classificadores de depressão, observa-se que apesar das medidas F serem relativamente semelhantes, cada função teve um efeito distinto nos resultados: enquanto Glorot aumentou o desempenho da métrica precisão, LeCun afetou principalmente a revocação. Um efeito diferente é observado para os classificadores de ansiedade e comorbidade, nos quais a função Glorot produziu um aumento da revocação, sem degradar a precisão. A análise estatística realizada ratifica essas observações, mostrando que a variação da função de inicialização do kernel produziu uma diferença significativa de performance para a métrica F1 em todos os classificadores fracos. Considerando os classificadores de comorbidade essa significância concentra-se na métrica de precisão. Já para os classificadores especialistas em uma condição alvo (ansiedade ou depressão), a métrica revocação foi significativamente afetada.

Figura B.1: Arquitetura LSTM: Impacto da Função de Inicialização do *Kernel*.



A Figura B.2 apresenta um comparativo entre os classificadores especialistas de arquitetura base CNN. Nela, observa-se que a variação da função de inicialização do kernel não produziu impacto significativo na performance dos classificadores de ansiedade. De modo oposto, os classificadores especializados em depressão apresentaram impacto significativo na performance em todas as métricas. Novamente, como observado nos modelos baseados na arquitetura LSTM, nesses classificadores a função LeCun beneficiou a revocação, enquanto a função Glorot beneficiou a precisão. Por fim, os classificadores de comorbidade apresentaram um aumento da precisão em detrimento da revocação ao usar a função Glorot. Embora essa variação entre revocação e precisão seja significativa, ela não representou um impacto significativo para a medida F1 desses classificadores.

Figura B.2: Arquitetura CNN: Impacto da Função de Inicialização do *Kernel*.



Os resultados desse experimento permitiram concluir que a variação do tipo de

função *kernel* para inicialização dos pesos da rede neural estabelece um *trade-off* entre revocação e precisão, sendo seu impacto maior para modelos baseados na arquitetura LSTM. Para os modelos de arquitetura CNN, verifica-se um impacto mínimo ou mesmo nulo na performance. De modo geral, não foi possível derivar um padrão de comportamento produzido pela variação dessa função quanto aos modelos gerados considerando a mesma arquitetura, mas funções diferentes. Por último, a análise comparativa entre arquiteturas destacou o comportamento similar observado na performance dos classificadores dedicados a identificação da depressão.

Neste trabalho, adotou-se o uso da função LeCun como uma variação para os modelos de arquitetura LSTM, explorando o comportamento observado nesse experimento como uma solução complementar para ganho de variabilidade do modelo *DAC Stacking*, quando considerando a função *meta-learning* do *Nível 1*.

APÊNDICE C — RELAÇÃO DE TERMOS MAIS FREQUENTES PARA OS TRANSTORNOS DE ANSIEDADE E DEPRESSÃO (DSM-5)

As tabelas C.1 e C.2 apresentam as listas de termos mais frequentes extraídos a partir da descrição para os Transtornos de Ansiedade e Depressão, respectivamente, segundo o manual DSM-5. As tabelas destacam em laranja os 7 comuns em ambos os transtornos. Em amarelo são destacados 4 termos de mesmo significado, porém com sintaxes diferentes.

Tabela C.1: Transtornos de Ansiedade: Relação de Termos conforme descrição DSM-5.

Lista de Termos Frequentes					
anxious	agoraphobia	apprehensive	arousal	attacks	avoid
avoidance	blank	clinging	control	crazy	crying
danger	depersonalization	difficult	discomfort	distress	escape
excessive	failure	falling	fear	feared	fearful
feelings	freezing	future	headache	help	humiliating
impairment	incapacitating	intense	interactions	irritability	losing
mind	mutism	nervios	neuroticism	palpitations	panic
paresthesias	persistent	phobia	restlessness	risk	screaming
shrinking	sleep	soreness	stress	tantrums	tension
thoughts	trembling	uncontrollable	unexpected	worry	

Tabela C.2: Transtornos Depressivos: Relação de Termos conforme descrição DSM-5.

Lista de Termos Frequentes					
aggression	angry	antidepressant	anxiety	appetite	behavioral
bereavement	bereavementrelated	capacity	careful	change	concentration
concerns	death	depressive	difficulty	disgust	displeasure
distress	disturbance	dyscontrol	dysphoric	dysregulation	dysthymia
empty	energy	explosive	extreme	facilitated	failure
fatigue	feelings	frustration	grief	guilt	hopelessness
hypersomnia	impairment	insomnia	interest	irritable	loss
mood	outbursts	overeating	pain	persistent	pleasure
retardation	sad	selfesteem	sleep	suffering	suicide
thoughts	timing	vulnerabilities	worse	worthlessness	

APÊNDICE D — ANÁLISE DE PERFORMANCE: CLASSIFICADORES ESPECIALISTAS (DETALHAMENTO)

No Experimento #1 (Seção 6.2.2), a análise de performance comparou o desempenho dos classificadores especialistas desenvolvidos usando as arquiteturas LSTM, CNN e Híbrida por condição alvo. Para essa análise, as comparações estatísticas foram realizadas considerando o resultado do experimento ao longo de 10 repetições. Esse experimento usou a parametrização base apresentada no Capítulo 5 para cada arquitetura, onde todos os modelos usam o mesmo *embedding* pré-treinado GloVe 6B. Os resultados em termos de Média (M) e Desvio Padrão (DP) para as repetições desse experimentos são apresentados na Tabela D.1. Cabe destacar que após essa análise estatística, para cada condição alvo, novos experimentos foram realizados para ajuste fino das arquiteturas exploradas resultando na performance apresentada nas Tabelas 6.1, 6.2 e 6.3 da Seção 6.2.2.

A Tabela D.1 permite verificar que, de modo geral, os modelos mantém uma performance regular mediante as 10 repetições, apresentando baixo desvio padrão para as métricas analisadas. Entre os modelos explorados, nota-se que de modo geral a arquitetura CNN é que apresenta a menor variação, em termos de desvio padrão, considerando as métricas analisadas para todas as condições alvo. Um comportamento inverso é observado para os modelos híbridos, em todas as condições alvo.

Tabela D.1: Experimento #1 - Classificadores Especialistas: Média e Desvio Padrão.

Condição Mental	**Arquitetura**	**Precisão**		**Revocação**		**F1**	
		M	**DP**	**M**	**DP**	**M**	**DP**
Ansiedade	LSTM	0,77	0,03	0,73	0,05	0,75	0,02
	CNN	0,71	0,00	0,80	0,00	0,75	0,00
	Híbrido	0,67	0,06	0,69	0,05	0,68	0,03
Depressão	LSTM	0,77	0,02	0,72	0,04	0,74	0,01
	CNN	0,82	0,04	0,68	0,07	0,74	0,02
	Híbrido	0,68	0,08	0,77	0,06	0,72	0,04
Comorbidade	LSTM	0,73	0,02	0,66	0,02	0,69	0,01
	CNN	0,82	0,00	0,70	0,01	0,76	0,01
	Híbrido	0,58	0,08	0,73	0,05	0,64	0,04

A Tabela D.2, apresenta o resultado do teste estatístico em termos de *p-value* para cada métrica analisada. Como mencionado na Seção 6.2.2, adotou-se o Teste Student T bicaudal pareado ($alpha = 0.5$) para estabelecer o comparativo de performance entre duas arquiteturas, considerando cada condição alvo. Cabe lembrar que, como hipótese nula tem-se que não há diferença entre duas arquiteturas comparadas para a tarefa de classificar uma determinada condição alvo. Como hipótese alternativa ($p - value < 0.05$), tem-se

Tabela D.2: Comparativos entre Arquiteturas: Resultado do Teste T Student.

Condição Mental	**Arquitetura**	***P-Value***		
		Precisão	**Revocação**	**F1**
Ansiedade	LSTM vs CNN	0,00	0,00	0,28
	LSTM vs Híbrido	0,00	0,17	0,00
	CNN vs Híbrido	0,09	0,00	0,00
Depressão	LSTM vs CNN	0,00	0,20	0,87
	LSTM vs Híbrido	0,01	0,04	0,05
	CNN vs Híbrido	0,00	0,01	0,08
Comorbidade	LSTM vs CNN	0,00	0,00	0,00
	LSTM vs Híbrido	0,00	0,00	0,00
	CNN vs Híbrido	0,00	0,12	0,00

que existe diferença de performance entre as arquiteturas analisadas. Assim, a Tabela D.2 apresenta o *p-value* resultante desse comparativo para as métricas Precisão, Revocação e medida F (F1) entre as combinações de arquitetura, por condição alvo.

Os resultados apresentados na Tabela D.1 mostram que todos os comparativos de performance realizados entre as arquiteturas registraram diferenças estatisticamente significativas para a tarefa de classificar uma mesma condição mental. Esses resultados permitem constatar que existe diferença de performance entre as diferentes arquiteturas exploradas, o que pode implicar em um ganho de variabilidade para o modelo de comitê proposto, à medida que as diferenças significativas observadas refletem um *trade-off* entre precisão e revocação ao considerar os comparativos entre duas arquiteturas.

APÊNDICE E — ANÁLISE DE PERFORMANCE: *DAC STACKING* E VARIAÇÕES (DETALHAMENTO)

No Experimento #3 (Seção 6.4), foram apresentadas diferentes análises comparativas de desempenho para os modelos *DAC Stacking* quanto à topologia (homogêneo vs heterogêneo) e às variações propostas (Base, EC e DT). As Tabelas E.1, E.2 e E.3 apresentam o valor médio das repetições resultantes do processo de treinamento por validação cruzada, juntamente com os valores de desvio padrão, para os modelos *DAC Stacking* e suas variações *EC* e *DT*, respectivamente.

Para confirmar se as diferenças de performance encontradas eram estatisticamente significativas, adotou-se o teste estatístico Student T bicaudal pareado ($alpha = 0,5$). Como hipótese nula ($p - value >= 0,05$), considerou-se que não há diferença de performance entre dois modelos comparados. Os resultados desse teste para os comparativos de performance entre os modelos *baseline* e os modelos *DAC Stacking* e variações *EC* e *DT* são apresentados nas Tabelas E.4, E.5 e E.6, respectivamente. Já as Tabelas E.7, E.8 e E.9 apresentam os resultados estatísticos para os comparativos de performance, considerando as variações de topologia (homogêneo vs heterogêneo) para os modelos *DAC Stacking*, *DAC Stacking EC* e *DAC Stacking DT*, respectivamente.

Por fim, a Tabela E.10 apresenta o comparativo de performance entre o melhor resultado, *DAC Stacking C*, e os modelos resultantes de cada variação proposta *DAC Stacking EC* e *DAC Stacking DT*.

Tabela E.1: Experimento #3 - *DAC Stacking*: Média (M) e Desvio Padrão (DP).

Modelo	**Multirrótulo**				**Controle**						**Ansiedade**						**Depressão**						**Comorbidade**					
DAC	**EMR**		**HL**		**Precisão**		**Revocação**		**F1**		**Precisão**		**Revocação**		**F1**		**Precisão**		**Revocação**		**F1**		**Precisão**		**Revocação**		**F1**	
Stacking	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP
L	0,46	0,00	0,29	0,00	0,67	0,01	0,77	0,01	0,72	0,00	0,62	0,01	0,69	0,03	0,65	0,01	0,57	0,00	0,66	0,02	0,61	0,01	0,33	0,00	0,72	0,02	0,45	0,00
C	0,42	0,01	0,28	0,00	0,74	0,02	0,63	0,03	0,68	0,01	0,62	0,02	0,73	0,06	0,67	0,02	0,58	0,02	0,72	0,04	0,64	0,01	0,33	0,01	0,75	0,02	0,46	0,01
H	0,42	0,02	0,29	0,00	0,75	0,01	0,64	0,05	0,69	0,03	0,60	0,02	0,74	0,01	0,67	0,01	0,57	0,01	0,67	0,01	0,61	0,01	0,33	0,00	0,76	0,00	0,46	0,00
LCH	0,40	0,02	0,30	0,02	0,75	0,01	0,59	0,07	0,66	0,04	0,58	0,04	0,81	0,04	0,67	0,01	0,56	0,01	0,71	0,06	0,63	0,01	0,32	0,02	0,79	0,05	0,46	0,01
Baseline L	0,33	0,04	0,37	0,03	0,57	0,05	0,63	0,03	0,60	0,04	0,57	0,01	0,39	0,14	0,45	0,10	0,55	0,01	0,35	0,16	0,41	0,11	0,29	0,01	0,37	0,14	0,32	0,04
Baseline C	0,28	0,12	0,33	0,03	0,73	0,06	0,48	0,28	0,53	0,24	0,62	0,01	0,36	0,08	0,45	0,06	0,62	0,02	0,37	0,13	0,46	0,09	0,33	0,01	0,36	0,09	0,34	0,04
Baseline H	0,31	0,04	0,36	0,01	0,60	0,03	0,52	0,13	0,54	0,05	0,54	0,04	0,47	0,16	0,49	0,08	0,55	0,03	0,46	0,09	0,50	0,04	0,30	0,02	0,44	0,11	0,35	0,02

Tabela E.2: Experimento #3 - *DAC Stacking EC*: Média (M) e Desvio Padrão (DP).

Modelo	**Multirrótulo**				**Controle**						**Ansiedade**						**Depressão**						**Comorbidade**					
DAC	**EMR**		**HL**		**Precisão**		**Revocação**		**F1**		**Precisão**		**Revocação**		**F1**		**Precisão**		**Revocação**		**F1**		**Precisão**		**Revocação**		**F1**	
Stacking EC	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP
L	0,46	0,01	0,29	0,00	0,67	0,01	0,78	0,02	0,72	0,00	0,61	0,01	0,69	0,02	0,65	0,00	0,57	0,01	0,64	0,04	0,60	0,02	0,32	0,01	0,71	0,03	0,45	0,01
C	0,42	0,01	0,28	0,01	0,74	0,01	0,66	0,06	0,69	0,03	0,63	0,02	0,68	0,09	0,65	0,04	0,58	0,02	0,71	0,05	0,64	0,01	0,33	0,01	0,72	0,09	0,45	0,01
H	0,43	0,02	0,29	0,01	0,72	0,03	0,68	0,06	0,70	0,02	0,60	0,01	0,76	0,04	0,67	0,01	0,56	0,02	0,69	0,05	0,62	0,01	0,32	0,02	0,76	0,04	0,44	0,01
LCH	0,41	0,02	0,29	0,02	0,75	0,01	0,61	0,05	0,67	0,03	0,62	0,04	0,73	0,10	0,66	0,02	0,57	0,04	0,69	0,11	0,62	0,02	0,33	0,03	0,75	0,11	0,46	0,01
Baseline L	0,33	0,04	0,37	0,03	0,57	0,05	0,63	0,03	0,60	0,04	0,57	0,01	0,39	0,14	0,45	0,10	0,55	0,01	0,35	0,16	0,41	0,11	0,29	0,01	0,37	0,14	0,32	0,04
Baseline C	0,28	0,12	0,33	0,03	0,73	0,06	0,48	0,28	0,53	0,24	0,62	0,01	0,36	0,08	0,45	0,06	0,62	0,02	0,37	0,13	0,46	0,09	0,33	0,01	0,36	0,09	0,34	0,04
Baseline H	0,31	0,04	0,36	0,01	0,60	0,03	0,52	0,13	0,54	0,05	0,54	0,04	0,47	0,16	0,49	0,08	0,55	0,03	0,46	0,09	0,50	0,04	0,30	0,02	0,44	0,11	0,35	0,02

Tabela E.3: Experimento #3 - *DAC Stacking DT*: Média (M) e Desvio Padrão (DP).

Modelo	**Multirrótulo**				**Controle**						**Ansiedade**						**Depressão**						**Comorbidade**					
DAC	**EMR**		**HL**		**Precisão**		**Revocação**		**F1**		**Precisão**		**Revocação**		**F1**		**Precisão**		**Revocação**		**F1**		**Precisão**		**Revocação**		**F1**	
Stacking DT	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP
L	0,46	0,00	0,29	0,00	0,68	0,01	0,77	0,01	0,72	0,00	0,61	0,01	0,71	0,02	0,66	0,01	0,57	0,00	0,66	0,01	0,61	0,00	0,32	0,00	0,73	0,02	0,45	0,00
C	0,42	0,02	0,28	0,01	0,73	0,01	0,65	0,05	0,69	0,03	0,62	0,02	0,72	0,07	0,67	0,02	0,57	0,01	0,72	0,03	0,63	0,01	0,33	0,01	0,75	0,05	0,46	0,01
LCH-LC	0,42	0,02	0,29	0,01	0,74	0,00	0,62	0,04	0,67	0,03	0,60	0,02	0,76	0,09	0,66	0,03	0,58	0,05	0,69	0,16	0,62	0,06	0,32	0,01	0,69	0,17	0,43	0,04
LCH-C	0,40	0,02	0,30	0,01	0,74	0,01	0,60	0,05	0,66	0,03	0,59	0,04	0,76	0,10	0,66	0,02	0,55	0,01	0,76	0,04	0,64	0,01	0,31	0,01	0,77	0,09	0,44	0,01
Baseline L	0,33	0,04	0,37	0,03	0,57	0,05	0,63	0,03	0,60	0,04	0,57	0,01	0,39	0,14	0,45	0,10	0,55	0,01	0,35	0,16	0,41	0,11	0,29	0,01	0,37	0,14	0,32	0,04
Baseline C	0,28	0,12	0,33	0,03	0,73	0,06	0,48	0,28	0,53	0,24	0,62	0,01	0,36	0,08	0,45	0,06	0,62	0,02	0,37	0,13	0,46	0,09	0,33	0,01	0,36	0,09	0,34	0,04
Baseline H	0,31	0,04	0,36	0,01	0,60	0,03	0,52	0,13	0,54	0,05	0,54	0,04	0,47	0,16	0,49	0,08	0,55	0,03	0,46	0,09	0,50	0,04	0,30	0,02	0,44	0,11	0,35	0,02

Tabela E.4: Experimento #3 -*DAC Stacking*: Comparativo estatístico com os *Baselines*.

Comparativo entre Modelos	**Teste Student T: *P-Value***													
	Multirrótulo		**Controle**			**Ansiedade**			**Depressão**			**Comorbidade**		
	(EMR)	**(HL)**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
DAC Stacking L vs Baseline L	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,01	0,01	0,00	0,00	0,00
DAC Stacking C vs Baseline C	0,05	0,03	0,94	0,27	0,23	0,69	0,00	0,00	0,01	0,00	0,01	0,23	0,00	0,00
DAC Stacking H vs Baseline H	0,01	0,00	0,00	0,06	0,00	0,01	0,02	0,01	0,28	0,01	0,00	0,02	0,00	0,00
DAC Stacking LCH vs Baseline L	0,02	0,02	0,00	0,23	0,06	0,51	0,00	0,01	0,09	0,00	0,01	0,00	0,00	0,00
DAC Stacking LCH vs Baseline C	0,09	0,05	0,51	0,42	0,27	0,08	0,00	0,00	0,00	0,01	0,02	0,33	0,00	0,00
DAC Stacking LCH vs Baseline H	0,02	0,00	0,00	0,25	0,02	0,21	0,01	0,01	0,49	0,01	0,00	0,17	0,00	0,00

Tabela E.5: Experimento #3 -*DAC Stacking EC*: Comparativo estatístico com os *Baselines*.

Comparativo entre Modelos	**Teste Student T: *P-Value***													
	Multirrótulo		**Controle**			**Ansiedade**			**Depressão**			**Comorbidade**		
	(EMR)	**(HL)**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
DAC Stacking EC L vs Baseline L	0,00	0,01	0,02	0,00	0,00	0,01	0,01	0,01	0,02	0,01	0,01	0,01	0,00	0,00
DAC Stacking EC C vs Baseline C	0,05	0,02	0,94	0,19	0,18	0,34	0,00	0,00	0,04	0,00	0,01	0,70	0,00	0,00
DAC Stacking EC H vs Baseline H	0,01	0,00	0,01	0,12	0,01	0,03	0,02	0,01	0,75	0,00	0,00	0,20	0,00	0,00
DAC Stacking EC LCH vs Baseline L	0,01	0,00	0,00	0,50	0,01	0,05	0,01	0,01	0,14	0,02	0,01	0,04	0,01	0,00
DAC Stacking EC LCH vs Baseline C	0,09	0,06	0,63	0,38	0,25	0,90	0,01	0,00	0,06	0,02	0,02	0,82	0,01	0,00
DAC Stacking EC LCH vs Baseline H	0,02	0,00	0,00	0,23	0,01	0,02	0,01	0,00	0,22	0,00	0,00	0,02	0,00	0,00

Tabela E.6: Experimento #3 -*DAC Stacking DT*: Comparativo estatístico com os *Baselines*.

Comparativo entre Modelos	**Teste Student T: *P-Value***													
	Multirrótulo		**Controle**			**Ansiedade**			**Depressão**			**Comorbidade**		
	EMR	**HL**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
DAC Stacking DT L vs Baseline L	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,02	0,02	0,00	0,01	0,00
DAC Stacking DT C vs Baseline C	0,06	0,03	0,86	0,28	0,23	0,78	0,00	0,00	0,00	0,01	0,01	0,62	0,00	0,00
DAC Stacking DT LCH-LC vs Baseline L	0,01	0,00	0,00	0,67	0,03	0,05	0,01	0,01	0,23	0,02	0,01	0,02	0,01	0,00
DAC Stacking DT LCH-LC vs Baseline C	0,06	0,09	0,79	0,34	0,25	0,19	0,00	0,00	0,22	0,00	0,01	0,28	0,01	0,01
DAC Stacking DT LCH-LC vs Baseline H	0,01	0,00	0,00	0,14	0,01	0,00	0,00	0,00	0,36	0,08	0,04	0,03	0,07	0,02
DAC Stacking DT LCH-C vs Baseline L	0,01	0,01	0,00	0,26	0,02	0,22	0,01	0,01	0,18	0,00	0,01	0,05	0,00	0,01
DAC Stacking DT LCH-C vs Baseline C	0,09	0,08	0,73	0,42	0,28	0,20	0,00	0,00	0,00	0,01	0,01	0,04	0,00	0,00
DAC Stacking DT LCH-C vs Baseline H	0,01	0,00	0,00	0,19	0,01	0,05	0,00	0,00	0,73	0,00	0,00	0,26	0,00	0,00

Tabela E.7: Experimento #3 - Comparativo estatístico entre variações *DAC Stacking*.

Comparativo	**Teste Student T: *P-Value***													
entre Modelos	**Multirrótulo**		**Controle**			**Ansiedade**			**Depressão**			**Comorbidade**		
DAC Stacking	**EMR**	**HL**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
L vs C	0,00	0,01	0,00	0,00	0,00	0,50	0,21	0,09	0,47	0,02	0,01	0,21	0,14	0,03
L vs H	0,00	0,43	0,00	0,01	0,08	0,11	0,02	0,11	0,18	0,18	1,00	0,37	0,02	0,00
L vs LCH	0,00	0,36	0,00	0,00	0,03	0,04	0,00	0,02	0,14	0,05	0,00	0,37	0,03	0,21
C vs H	0,38	0,11	0,12	0,65	0,22	0,08	0,77	0,50	0,30	0,03	0,00	0,37	0,37	1,00
C vs LCH	0,06	0,11	0,02	0,18	0,32	0,09	0,08	0,85	0,08	0,54	0,14	0,23	0,11	0,59
H vs LCH	0,21	0,29	1,00	0,12	0,10	0,28	0,02	0,37	0,21	0,14	0,07	0,27	0,23	0,37

Tabela E.8: Experimento #3 - Comparativo estatístico entre variações *DAC Stacking EC*.

Comparativo	**Teste Student T: *P-Value***													
entre Modelos	**Multirrótulo**		**Controle**			**Ansiedade**			**Depressão**			**Comorbidade**		
DAC Stacking EC	**EMR**	**HL**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
L vs C	0,01	0,01	0,00	0,01	0,11	0,02	0,71	0,90	0,34	0,01	0,00	0,19	0,79	0,47
L vs H	0,06	0,82	0,02	0,03	0,13	0,34	0,03	0,01	0,20	0,17	0,24	0,59	0,15	0,70
L vs LCH	0,00	0,57	0,00	0,01	0,02	0,61	0,53	0,28	0,92	0,40	0,29	0,58	0,55	0,23
C vs H	0,45	0,09	0,51	0,69	0,74	0,06	0,17	0,37	0,22	0,59	0,03	0,29	0,45	0,46
C vs LCH	0,27	0,47	0,24	0,40	0,50	0,61	0,56	0,57	0,82	0,76	0,24	1,00	0,75	0,18
H vs LCH	0,00	0,59	0,14	0,05	0,06	0,39	0,57	0,69	0,21	0,96	0,50	0,33	0,78	0,24

Tabela E.9: Experimento #3 - Comparativo estatístico entre variações *DAC Stacking DT*.

Comparativo	**Teste Student T: *P-Value***													
entre Modelos	**Multirrótulo**		**Controle**			**Ansiedade**			**Depressão**			**Comorbidade**		
DAC Stacking DT	**EMR**	**HL**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
L vs C	0,01	0,16	0,00	0,00	0,03	0,28	0,84	0,23	1,00	0,00	0,00	0,16	0,35	0,07
L vs LCH-LC	0,02	0,97	0,00	0,00	0,03	0,24	0,28	0,55	0,68	0,69	0,84	0,54	0,61	0,47
L vs LCH-C	0,00	0,20	0,00	0,00	0,01	0,26	0,29	0,51	0,02	0,01	0,01	0,14	0,36	0,18
C vs LCH-LC	0,90	0,40	0,11	0,45	0,60	0,17	0,39	0,87	0,71	0,76	0,58	0,46	0,54	0,19
C vs LCH-C	0,00	0,01	0,03	0,00	0,00	0,03	0,32	0,89	0,05	0,12	0,48	0,01	0,62	0,01
LCH-LC vs LCH-C	0,39	0,36	0,62	0,65	0,66	0,62	0,76	1,00	0,31	0,42	0,50	0,16	0,39	0,76

Tabela E.10: Experimento #3 - Comparativo estatístico entre *DAC Stacking C* e Variações.

Comparativo entre	***Teste Student T: P-Value***													
DAC Stacking C (C) e	**Multirrótulo**		**Controle**			**Ansiedade**			**Depressão**			**Comorbidade**		
Variações (EC e DT)	**EMR**	**HL**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
C vs EC L	0,01	0,00	0,00	0,00	0,00	0,02	0,23	0,04	0,62	0,01	0,01	0,03	0,01	0,00
C vs EC C	0,79	0,47	1,00	0,43	0,35	0,37	0,12	0,12	0,79	0,78	0,75	0,70	0,42	0,29
C vs EC H	0,46	0,06	0,54	0,27	0,15	0,04	0,44	0,75	0,06	0,07	0,00	0,15	0,69	0,08
C vs EC LCH	0,18	0,51	0,07	0,49	0,88	0,81	1,00	0,33	0,88	0,54	0,24	0,90	0,94	0,59
C vs DT L	0,00	0,02	0,00	0,00	0,00	0,14	0,63	0,20	0,62	0,03	0,01	0,02	0,28	0,00
C vs DT C	0,78	0,72	0,43	0,45	0,39	1,00	0,83	0,63	0,69	0,76	0,37	1,00	0,94	1,00
C vs DT LCH-LC	0,80	0,11	0,47	0,60	0,89	0,14	0,50	0,48	0,80	0,64	0,44	0,03	0,43	0,20
C vs DT LCH-C	0,02	0,05	0,18	0,15	0,28	0,19	0,53	0,54	0,00	0,20	1,00	0,03	0,54	0,03

APÊNDICE F — ANÁLISE DE PERFORMANCE: BERT E VARIAÇÕES VS DAC STACKING (DETALHAMENTO)

No Experimento #5 (Seção 6.6) apresentou-se uma análise comparativa de performance entre os modelos BERT MP (classificador principal) e suas variações usando BERT como *embedding* pré-treinado para as arquiteturas LSTM (BERT L), CNN (BERT C) e Híbrida (BERT H). A Tabela F.1 apresenta os resultados da média de repetições, juntamente com o desvio padrão para cada variação testada.

Assim como os demais experimentos, para confirmar a significância estatística das diferenças observadas considerou-se o Test Student T, configurado conforme descrito na Seção 6.6. A Tabela F.2 apresenta o resultado, em termos de *p-value* do teste estatístico comparando BERT e suas variações. A última linha dessa tabela apresenta um comparativo entre os modelos de melhor performance BERT H e *DAC Stacking DT*.

Tabela F.1: *Modelos BERT*: Média (M) e Desvio Padrão (DP).

Modelo	Multirrótulo				Control						Ansiedade						Depressão						Comorbidade					
	EMR		HL		Precisão		Revocação		F1		Precisão		Revocação		F1		Precisão		Revocação		F1		Precisão		Revocação		F1	
	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP
BERT MP	0,42	0,02	0,28	0,01	0,72	0,04	0,73	0,08	0,72	0,02	0,64	0,01	0,61	0,04	0,63	0,01	0,62	0,02	0,50	0,14	0,54	0,09	0,34	0,02	0,54	0,18	0,41	0,07
BERT L	0,45	0,01	0,28	0,01	0,73	0,02	0,70	0,04	0,72	0,01	0,62	0,01	0,72	0,02	0,66	0,01	0,57	0,02	0,76	0,04	0,65	0,02	0,32	0,02	0,75	0,03	0,45	0,01
BERT C	0,03	0,00	0,39	0,00	0,83	0,06	0,05	0,01	0,10	0,01	0,78	0,22	0,04	0,02	0,07	0,04	0,64	0,18	0,04	0,02	0,07	0,03	0,37	0,25	0,04	0,03	0,07	0,06
BERT H	0,44	0,02	0,28	0,01	0,75	0,02	0,67	0,05	0,71	0,02	0,61	0,01	0,74	0,04	0,67	0,01	0,56	0,01	0,77	0,05	0,65	0,01	0,32	0,01	0,77	0,03	0,45	0,01
DAC Stacking DT	0,42	0,01	0,28	0,00	0,74	0,02	0,63	0,03	0,68	0,01	0,62	0,02	0,73	0,06	0,67	0,02	0,58	0,02	0,72	0,04	0,64	0,01	0,33	0,01	0,75	0,02	0,46	0,01

Tabela F.2: *Modelos BERT vs DAC Stacking DT*: Comparativo Estatístico de Performance.

Comparativo entre Modelos BERT e DAC Stacking DT	*Teste Student T: P-Value*														
	Multirrótulo		Controle			Ansiedade			Depressão			Comorbidade			
	EMR	HL	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
MP vs BERT L	0,13	0,92	0,66	0,55	0,67	0,04	0,00	0,00	0,00	0,01	0,04	0,21	0,05	0,26	
MP vs BERT C	0,00	0,00	0,01	0,00	0,00	0,22	0,00	0,00	0,88	0,00	0,00	0,78	0,00	0,00	
MP vs BERT H	0,10	0,53	0,22	0,21	0,40	0,01	0,00	0,00	0,00	0,01	0,05	0,15	0,03	0,32	
BERT L vs BERT C	0,00	0,00	0,04	0,00	0,00	0,15	0,00	0,00	0,49	0,00	0,00	0,69	0,00	0,00	
BERT L vs BERT H	0,50	0,36	0,03	0,13	0,40	0,47	0,10	0,25	0,00	0,50	0,73	0,30	0,22	0,74	
BERT C vs BERT H	0,00	0,00	0,09	0,00	0,00	0,15	0,00	0,00	0,40	0,00	0,00	0,65	0,00	0,00	
BERT H vs DAC Stacking DT	0,17	0,93	0,38	0,26	0,07	0,28	0,77	0,88	0,08	0,17	0,28	0,05	0,54	0,00	