UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO

DETECÇÃO AUTOMÁTICA DE POSTAGENS POSSIVELMENTE DEPRESSIVAS EM REDES SOCIAIS

Augusto Rozendo Mendes

Orientadora: Profa. Dra. Helena de Medeiros Caseli

São Carlos - SP

22 de novembro de 2021

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO

DETECÇÃO AUTOMÁTICA DE POSTAGENS POSSIVELMENTE DEPRESSIVAS EM REDES SOCIAIS

Augusto Rozendo Mendes

Monografia apresentada ao Curso de Graduação em Ciência da Computação da Universidade Federal de São Carlos, para a obtenção do título de bacharel em Ciência da Computação.
Orientadora: Profa. Dra. Helena de Medeiros Caseli

São Carlos - SP
22 de novembro de 2021

AGRADECIMENTOS

Ao Prof. Dr. Ivandré Paraboni (EACH/USP) pela disponibilização do *corpus* de (SANTOS; FUNABASHI; PARABONI, 2020) utilizado nos experimentos apresentados neste trabalho.

Ao Dr. Amir Yazdavar por compartilhar o léxico de depressão original, em inglês, em <https://github.com/yazdavar/Social-media-Depression-Detector/blob/master/depression_lexicon.json>, o qual foi utilizado neste trabalho.

Ao CNPq e à FAPESP pelo apoio financeiro, uma vez que este trabalho foi desenvolvido com o apoio de bolsa PIBIC (entre setembro de 2020 e setembro de 2021) e faz parte do Projeto AMIVE (Auxílio RegularFAPESP #20/05157-9).

RESUMO

A depressão é uma das questões de saúde mental mais preocupantes da atualidade. No Brasil, em 2019, 10,2% da população adulta relatou ter sido diagnosticada com depressão segundo dados da Pesquisa Nacional de Saúde. Identificar pessoas com perfil possivelmente depressivo permite um acompanhamento adequado por parte dos profissionais de saúde mental. Nesse sentido, as redes sociais online, como o Twitter, podem ser importantes aliadas. Esta monografia apresenta experimentos realizados para a classificação automática de postagens individuais (ou conjuntos de postagens de um mesmo usuário) do Twitter contendo conteúdo que denota algum sintoma de depressão, bem como classificação de postagens depressivas e usuários deprimidos por meio de modelo *ensemble* composto por classificadores de sintomas. A classificação com regressão logística apresentou os melhores resultados em ambas as tarefas de classificação de sintomas e depressão ($F1$ média de 57% no caso da primeira, $F1$ de 64% no caso da segunda). Este trabalho é parte do projeto Amive (Auxílio Regular FAPESP $\#20/05157\text{-}9$).

Palavras-chave: depressão, Twitter, PHQ-9, saúde mental, aprendizado supervisionado

ABSTRACT

Currently, depression is one of the most worrisome mental health issues. In Brazil, in 2019, 10.2% of the adult population reported having been diagnosed with depression according to data from the National Health Survey. Identifying people with a possible depressive profile allows adequate monitoring by mental health professionals. In this sense, online social networks such as Twitter can be important allies. This monography presents experiments carried out for the automatic classification of Twitter posts (or a collection of posts produced by a given user) containing content that denotes some symptom of depression, as well as classification of depressive posts and users through an ensemble model composed of symptom classifiers. Logistic regression showed the best results in both symptom and depression classification tasks (average $F1$ equal to 57% for the former, $F1$ equal to 64% for the latter). This work is part of the Amive project (FAPESP Regular Grant $\#20/05157$-9).

Keywords: depression, Twitter, PHQ-9, mental health, supervised learning

SUMÁRIO

1 INTRODUÇÃO

O tratamento de depressão pode ser considerado uma das mais importantes questões de saúde mental da atualidade. No Brasil, segundo dados da Pesquisa Nacional de Saúde realizada em 2019, 10,2% dos adultos brasileiros mencionaram terem recebido diagnóstico de depressão por profissional de saúde mental[1].

A depressão maior pode ser diagnosticada quando uma pessoa apresenta, por pelo menos duas semanas, cinco ou mais sintomas, sendo que pelo menos um dos sintomas é humor deprimido ou perda de interesse ou prazer; os outros sintomas compreendem: diminuição ou aumento de peso ou apetite; insônia ou hipersonia; agitação ou retardo psicomotor; fadiga ou perda de energia; sentimentos de inutilidade ou culpa; concentração diminuída, ou indecisão e/ou pensamentos recorrentes de morte. Segundo o DSM-5 (American Psychiatric Association et al., 2014), esses sintomas devem causar sofrimento clinicamente significativo ou prejuízo no funcionamento social, profissional ou em outras áreas importantes da vida do indivíduo.

Identificar sinais de depressão possibilita intervenções e acompanhamento por profissionais de saúde mental. Além disso, em um nível macroscópico, a identificação de padrões de depressão em comunidades facilita a compreensão acerca do fenômeno, abrindo caminho para políticas de saúde pública.

Atualmente, uma das formas de rastreamento de sintomas de depressão ao longo do tempo são instrumentos como o PHQ-9 (KROENKE; SPITZER; WILLIAMS, 2001) e escala Hamilton, questionários aplicados por profissionais médicos. Segundo (COPPERSMITH et al., 2018), este método apresenta limitações em termos de alcance (o entrevistado deve buscar por ajuda médica) e depende da lembrança do estado emocional ao longo do tempo por parte do paciente. Uma possível forma de complementar metodologias vigentes de identificação de depressão é por meio da análise de conteúdo produzido por usuários de redes sociais, que pode conter pistas sobre o estado emocional dos mesmos.

Dada a ubiquidade de plataformas como Facebook e Twitter, é possível que esta análise abranja um número maior de pessoas, contribuindo para o entendimento da depressão em nível populacional, além de possivelmente auxiliar no tratamento de casos previamente não identificados. Nesta frente, destacam-se estudos voltados para a classificação automática de postagens extraídas de mídias sociais na área de saúde mental. Em sua maioria, estes estudos fazem uso apenas de conteúdo textual para o treinamento de modelos preditivos, tratando-se de aplicação pura de Processamento de Linguagem Natural (PLN) (JI et al., 2018; COPPERSMITH et al., 2018), embora alguns estudos, como (CHOUDHURY; COUNTS;

[1] Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=29270&t=resultados>

HORVITZ, 2013) incorporem metadados como horário de postagem e métricas de engajamento de uma dada rede (como número de *retweets* no Twitter), e também tenha sido investigada a construção de modelos multimodais que fazem uso de texto e imagem (MANN; PAES; MATSUSHIMA, 2020).

Em linhas gerais, estes trabalhos tratam da detecção de postagens possivelmente depressivas (PPDs), e usam os modelos treinados para a classificação de postagens individuais, de usuários, ou para medir o nível de depressão em populações. Este interesse na aplicação de técnicas de PLN para saúde mental é fruto da alta disponibilidade de dados proporcionada pelas Redes Sociais Online (RSO) por meio de *Application Programming Interfaces* (APIs). No entanto, a conexão entre as *features* extraídas/modelos gerados e a literatura médica é um tema pouco explorado. A pesquisa descrita nesta monografia buscou aproximar a classificação de PPDs com práticas médicas vigentes, por meio de um modelo baseado na presença de sintomas amplamente usados para diagnóstico de depressão (instrumento de 9 questões Patient Health Questionnaire, PHQ-9) .

Muitos dos esforços em PLN são voltados ao processamento na língua inglesa, o que é refletido na relativa escassez de estudos acerca do tópico em língua portuguesa. De particular interesse entre estes é o elaborado por (SANTOS; FUNABASHI; PARABONI, 2020), visto que o *corpus* produzido durante o mesmo foi usado nesta pesquisa.

Tendo em mente os trabalhos anteriores, este trabalho teve como objetivo **treinar classificadores automáticos de PPDs em RSO, escritas no português do Brasil, com base na presença de sintomas de depressão**. Para tanto, foram definidas *features*, aplicadas ferramentas de PLN e algoritmos de Aprendizado de máquina (AM) para treinar modelos e avaliar sua capacidade em distinguir PPD. O modelo proposto é composto por 9 classificadores automáticos, sendo que cada um detecta a presença de um dado sintoma avaliado pelo PHQ-9 em uma postagem. Desta forma, buscou-se construir um modelo com maior capacidade explicativa, pois seria possível apontar quais indícios de sintomas presentes nas sentenças levaram à classificação de uma postagem como PPD.

Estes esforços resultaram nas seguintes contribuições científicas:

1. **Léxico de termos relativos aos 9 sintomas de depressão** avaliados pelo PHQ-9, em português do Brasil, gerados com base no léxico elaborado por Yazdavar et al. (2017) para o inglês.

2. **Classificadores capazes de identificar estes sintomas** com F1 (média harmônica entre precisão e revocação) entre $54,7\%$ e $60,2\%$, tratando-se, até onde sabemos, de **primeira avaliação da tarefa de classificação de sintomas de depressão no português**.

3. ***Word embeddings*** treinadas a partir do *corpus* usado neste trabalho e, portanto,

específicas do domínio de saúde mental em redes sociais.

Este trabalho está dividido em 5 capítulos, incluindo esta introdução. Na seção 2 são descritos alguns dos trabalhos relacionados supracitados. A seção 3 descreve o modelo proposto, recursos utilizados e análise exploratória dos mesmos. A seção 4 traz resultados de performance dos modelos gerados, interpretação desses resultados e análise qualitativa de explicabilidade do modelo. Por fim, a seção 5 apresenta conclusões resultantes dos experimentos, bem como aspectos a serem explorados por trabalhos futuros.

2 TRABALHOS RELACIONADOS

Este capítulo descreve os trabalhos mais relacionados à pesquisa desenvolvida neste trabalho de conclusão de curso, e a Tabela 1 resume as principais características dos trabalhos aqui apresentados.

Ji et al. (2018) fizeram uma análise exploratória de postagens extraídas das plataformas de mídia social Reddit e Twitter, com diferentes métodos de coleta. No caso do Reddit, as postagens da classe positiva foram coletadas da comunidade de apoio mútuo para indivíduos com ideação suicida r/SuicideWatch, e as postagens da classe negativa foram coletadas de comunidades sem relação com suicídio, como r/books. Já no caso do Twitter, foram buscadas postagens que continham palavras-chave relacionadas com ideação suicida, que foram posteriormente anotadas.

A partir destes *corpora* resultaram observações sobre o estilo de linguagem de postagens indicativas de ideação suicida, como o uso maior de pronomes pessoais e uso maior dos tempos presente e futuro, além de diferenças no conteúdo: postagens do Reddit costumam relatar as experiências do autor com maior detalhe, já postagens do Twitter tendem a ser mais diretas e agressivas, como pode ser observado na nuvem de palavras da Figura 1. Essas observações demonstram a utilidade de *features* capazes de capturar estilo linguístico para a análise de ideação suicida em texto, e a influência que a escolha de plataforma pode ter sobre o conteúdo coletado.

Figura 1 – Nuvem de palavras de postagens do Reddit (esquerda) e Twitter (direita)

Fonte: Ji et al. (2018)

Após a análise exploratória, foram treinados modelos com diferentes combinações de *features*, *corpora* e algoritmos de aprendizado de máquina. Os *features* visavam capturar conteúdo estatístico (número de palavras, *tokens* e caracteres), sintático (*part-of-speech*[1]) e linguístico (*Linguistic Inquiry and Word Count*, LIWC), além de frequência de palavras (*Term-Frequency Inverted Document Frequency*, TF-IDF), tópicos (extraídos por *Latent Dirichlet Allocation*, LDA) e, para o treinamento de modelos de aprendizado profundo, *word embeddings*.

[1] Categoria morfossintática da palavra como verbo, substantivo, adjetivo, advérbio, etc.

Foram avaliados modelos gerados pelos seguintes algoritmos de AM tradicionais: Support Vector Machine (SVM), Random Forest, Árvores de Decisão com Gradient Boosting (GBDT), XGBoost (variação de GBDT) e redes neurais *feed-forward* de múltiplas camadas (MLFFNN), bem como a técnica de aprendizado profundo Long Short-Term Memory (LSTM). Todos os modelos treinados apresentaram boa performance (F1-score acima de 0,75) independentemente do *corpus* ou algoritmo usado. Observou-se que a performance dos modelos aumentou conforme mais *features* eram usados (melhor modelo apresentou F1-score de 0,95), o que sugere que todos capturaram informação relevante.

Choudhury, Counts e Horvitz (2013) investigaram o uso de classificadores automáticos de postaganes depressivas como forma de análise de depressão em populações, visto que avaliações de depressão em nível populacional são custosos e relativamente infrequentes. O estudo propôs a métrica *Social Media Depression Index* (SMDI) para este fim, sendo esta calculada a partir das frequências de postagens depressivas e não depressivas em uma dada região e dia.

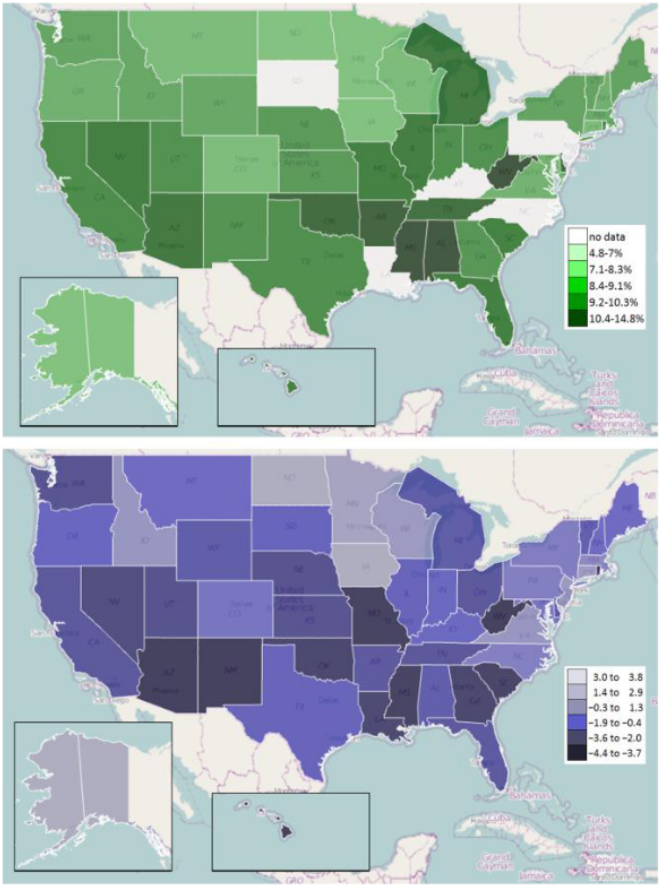
O *corpus* de postagens usado para o treinamento do modelo foi resultado de uma tarefa de *crowdsourcing* usando a plataforma Mechanical Turk, em que participantes foram instruídos a responder: (i) um questionário para diagnóstico de depressão, (ii) se tinham histórico de depressão, e (iii) podiam optar por compartilhar seu nome de usuário na plataforma Twitter para que suas postagens fossem coletadas.

Para o modelo responsável pela classificação de postagens, foram usados *features* que capturam estilo linguístico (LIWC), emoção (LIWC), frequência de termos (unigramas e bigramas) e horário da postagem (dividido entre dia e noite), além de dados de engajamento (número de postagens, respostas, compartilhamentos,etc.) e de *ego-network* (número de seguidores e seguidos). Dada a alta dimensionalidade do espaço de *features*, foi aplicada técnica de redução de dimensionalidade *Principal Component Analysis* (PCA), resultando em aumento de performance. O modelo, treinado por Support Vector Machine (SVM) atingiu acurácia de 74,57%.

Demonstrou-se que, de fato, a classificação de depressão em postagens individuais é confiável o suficiente para gerar informação sobre populações, pois os valores de SMDI calculados apresentaram correlação positiva significativa com dados coletados pelo *Center for Disease Control* (CDC) relativos a populações, tanto em termos de localidade (como ilustra a Figura 2), quanto gênero.

Uma tarefa similar à detecção de depressão/risco de suicídio em nível de postagem é esta mesma detecção em nível de usuário. Em experimento conduzido por Eichstaedt et al. (2018), indivíduos que apresentavam histórico de depressão em seu *Eletronic Medical Record* (EMR) concederam acesso à suas postagens na plataforma Facebook, a partir das quais foram criados 7 *corpora* compostos por postagens anteriores ao diagnóstico. Cada *corpus* continha textos postados em intervalos de tempo distintos, de forma a avaliar a capacidade de predição

Figura 2 – Comparação visual entre estatísticas do CDC (imagem superior) e métrica SMDI (inferior).



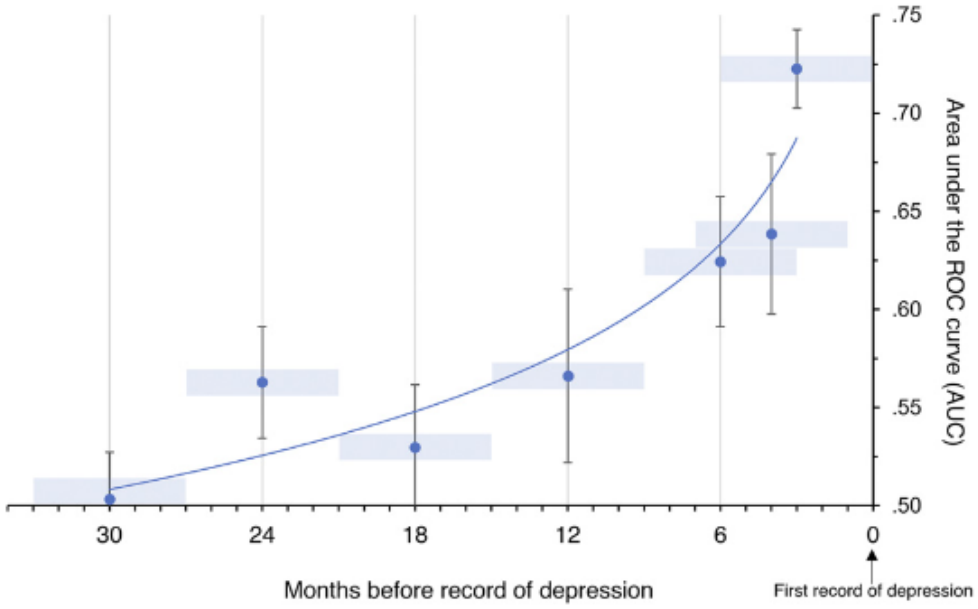
Fonte: Choudhury, Counts e Horvitz (2013)

do modelo conforme a distância da data do diagnóstico. Observou-se que a performance do modelo diminui drasticamente conforme a data das postagens se distancia da data do diagnóstico, com predições em intervalos de tempo acima de 9 meses sendo consideradas inadequadas. A partir desta constatação, os autores tomaram como limiar de boa discriminação o valor de AUC[2] igual a 0,7, sendo valores abaixo deste considerados inadequados. A Figura 3 traz os resultados em detalhe.

Segundo a literatura médica (American Psychiatric Association et al., 2014), uma das características associadas à depressão é a ruminação, que é a ocorrência de pensamentos repetitivos associados a eventos passados. Nambisan et al. (2015) investigaram se usuários de redes sociais com depressão demonstram ruminação em suas postagens. Para isso, foram coletadas postagens do Twitter que continham palavras-chave relacionadas a auto-declaração de depressão. Verificou-se manualmente se esses textos eram de fato auto-declarativos, e em caso positivo todas as postagens do usuário foram coletadas. Criou-se também um grupo de controle, composto por postagens de usuários que não produziram textos contendo as

[2] AUC é a área abaixo de uma curva *Receiver Operating Characteristics*, que expressa a relação entre a taxa de Falsos Positivos e Positivos Reais com a variação de valor limiar para predição positiva.

Figura 3 – Variação de AUC de acordo com o afastamento da data do primeiro indício de depressão



Fonte: Eichstaedt et al. (2018)

palavras-chave usadas para a coleta da classe positiva.

Foram investigadas postagens acerca de dois sintomas somáticos de depressão – variação no sono e dor – e ideação suicida. Foi constatado, por meio do teste exato de Fisher para avaliação de relevância estatística e teste chi2 para avaliação de correlação, que usuários deprimidos postam sobre os sintomas somáticos e ideação suicida com maior frequência que indivíduos não deprimidos, e que estas postagens compõem parte maior do total de postagens, o que indica ruminação acerca destes tópicos.

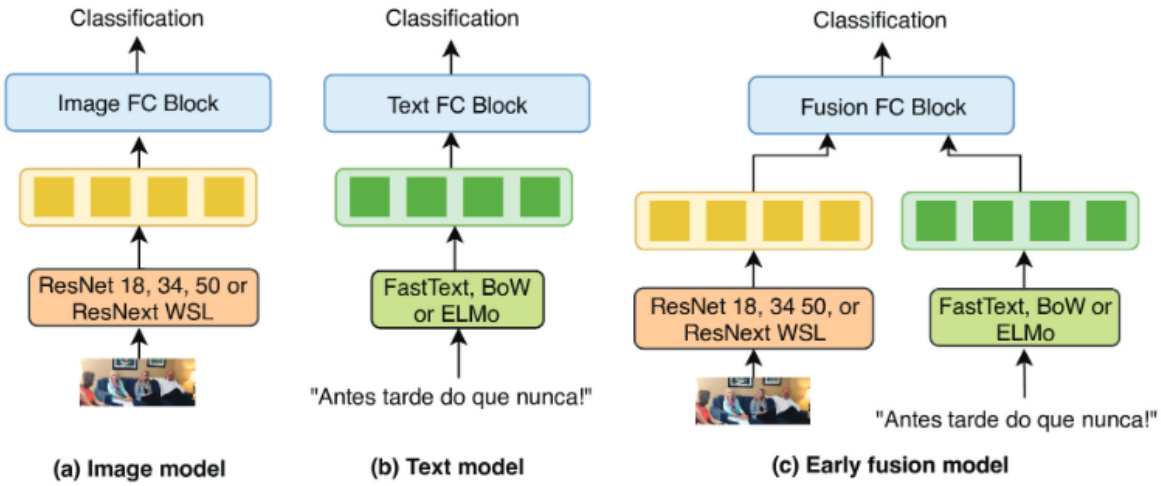
Santos, Funabashi e Paraboni (2020) descrevem a construção de um *corpus* de postagens de usuários brasileiros do Twitter, com o intuito de facilitar o avanço de pesquisas de tarefas relativas à saúde mental para a língua portuguesa. Essas postagens foram coletadas de indivíduos que auto-declararam diagnósticos de saúde mental – como depressão, ansiedade, etc. – em alguma postagem (buscou-se por expressões como "Eu + diagnosticado + depressão") para a classe positiva e de indivíduos que manifestam interesse por tópicos de saúde mental por outros motivos (como engajamento com a campanha de conscientização Setembro Amarelo). Todas as postagens usadas para identificar os usuários como pertencentes a uma dada classe foram avaliadas manualmente de forma a averiguar que de fato pertencem a uma dada classe, sendo que no caso das postagens da classe positiva era preciso haver a especificação de data de diagnóstico.

Foram também conduzidos 3 experimentos. Dois destes trataram da tarefa de classifica-

ção de postagens individuais, sendo que em um deles só foram usadas postagens que continham o termo "Eu" (com o intuito de incluir apenas mensagens que tratam das experiências subjetivas do usuário). O último experimento tratou da detecção de problemas de saúde mental em nível de usuário com diferentes volumes de postagens antecedentes ao diagnóstico. Em todos os casos, o modelo que apresentou melhor performance foi a Regressão Logística com TF-IDF, alcançando $F1$ de 69% para a classificação de postagens, Regressão $F1$ de 64% para a classificação de postagens que continham o termo "Eu", e $F1$ de 64% para a classificação de usuários usando as 500 primeiras postagens.

Outro estudo voltado para o português que se destaca foi realizado por Mann, Paes e Matsushima (2020) no qual foi proposto um modelo multimodal para classificação de postagens depressivas. Tal modelo se baseia tanto no conteúdo textual das postagens quanto nas imagens. Visto que a tarefa era de treinamento de modelo multimodal, a rede social escolhida para a coleta dos dados foi o Instagram, pois tem como característica a ênfase no compartilhamento de imagens. O modelo proposto concatena *features* textuais (representações geradas por modelo ELMo) e visuais (geradas por modelos Resnet e Resnext) através de camada de fusão totalmente conectada (ilustrada na Figura 4). A partir destas *features*, uma rede neural foi treinada para a classificação de postagens com indícios de depressão.

Figura 4 – Modelos de classificação baseados em imagem, texto e multimodal propostos por Mann, Paes e Matsushima (2020).



Fonte: Mann, Paes e Matsushima (2020)

Para fins de comparação, foram investigados também modelos que utilizam apenas *features* textuais ou visuais, além do uso de engenharia de *features* tradicional, tanto para imagem quanto para texto. Dentre os modelos resultantes, o modelo multimodal baseado em aprendizado profundo alcançou melhor performance (F1 de 79%). Também observou-se que *features* textuais levaram a resultados melhores do que as visuais, e modelos baseados em engenharia de *features* tradicionais, apesar de competitivos, sofrem perda de performance quando usam *features* visuais e textuais concatenadas.

Yazdavar et al. (2017) investigaram o uso de um modelo semi-supervisionado para a detecção de sintomas de depressão, com o intuito de facilitar o uso de grandes volumes de dados. Foi proposta a classificação de conjuntos de postagens produzidas em intervalos de tempo fixos, de forma a produzir uma série temporal, alinhando a tarefa com a literatura médica, pois a duração e frequência de episódios depressivos é crucial para o diagnóstico de depressão maior.

O modelo proposto (ssToT) introduz conhecimento de domínio na modelagem de tópicos através de um léxico contendo termos relevantes para cada sintoma, um subconjunto dos quais são designados a tópicos distintos, de forma a guiar o modelo para a tarefa proposta. Este modelo foi avaliado qualitativamente, e então comparado com *baselines* não-supervisionados e semi-supervisionados (visto que o modelo gera tópicos), e com classificadores multirrótulo (visto que pode-se usar o modelo com classificador, por meio da distribuição de tópicos em uma dada janela de tempo). Na tarefa de classificação, o modelo proposto obteve performance melhor que os *baselines* (Naive Bayes multinomial e SVM, com técnicas de classificação multirrótulo Cadeia de Classificadores e Relevância Binária, totalizando 4 *baselines*) em 5 dos 9 sintomas, e apresentou melhor acurácia média.

Yadav et al. (2020) investigaram a classificação de sintomas em postagens do Twitter por meio de aprendizado multitarefa, técnica na qual dois modelos são treinados (ajuste fino de modelo BERT pré treinado) em paralelo para tarefas distintas, mas relacionadas, contribuindo para o ajuste de parâmetros um do outro. Os autores propuseram uma nova técnica de aprendizado multitarefa denominada *Figurative Language enabled MultiTask Learning*, que realiza a troca de parâmetros por meio da combinação das representações de palavras (*tokens*) geradas por cada modelo (combinação linear), a qual tem seus coeficientes ajustados durante treinamento. Ou seja, as representações de palavras geradas por cada modelo são combinadas, efetivamente gerando uma nova representação contendo informação das duas tarefas.

No caso, a tarefa primária escolhida foi a classificação de sintomas e a secundária a classificação de linguagem figurada, visto que um obstáculo recorrente para a classificação de postagens relacionadas ao tema de depressão, segundo a literatura, é o uso ubíquo de figuras de linguagem como ironia e sarcasmo em postagens, pois os termos contidos nestas não apresentam significado usual.

Para o treinamento, foram coletados 12.155 *tweets*, sendo que 3.738 destes foram coletados de indivíduos auto-declarados deprimidos, encontrados por meio de busca por palavras-chave contidas no léxico de Yazdavar et al. (2017) supracitado. Esses *tweets* foram então anotados manualmente com rótulos para as duas tarefas. No caso da tarefa de detecção de sintomas, foram usados os 9 sintomas avaliados pelo PHQ-9 como rótulos, e no caso da tarefa de detecção de linguagem figurada, utilizou-se os rótulos "metáfora", "sarcasmo" e "outros".

Após os treinamentos dos modelos, estes foram comparados com 4 *baselines*, sendo 1 destes ajuste fino regular (STL-BERT) e os demais multitarefa: MTL-H-BERT (denso),

MTL-S-BERT (*Cross-Stitch*) e MTL-S-BERT (Co-Attention). A técnica proposta superou os *baselines* nas duas tarefas, com $F1$ de $75,03\%$ para a tarefa de classificação de sintomas, e $F1$ de $75,55\%$ para a tarefa de classificação de linguagem figurada.

Tabela 1 – Visão geral dos trabalhos discutidos

Autores	**Tarefa(s)**
Ji et al. (2018)	Classificação binária de ideação suicida em postagens do Twitter e Reddit, com amplo conjunto de *features* e técnicas
Choudhury, Counts e Horvitz (2013)	Classificação de postagens depressivas do Twitter com base em *features* morfossintáticas, emocionais, horário de postagem, e baseadas em engajamento na rede social, utilizando SVM. O melhor modelo foi então usado para identificar nível de depressão em escala populacional.
Eichstaedt et al. (2018)	Classificação de usuários do Facebook (deprimido/ não deprimido) por modelo de Regressão Logística treinado com *features* de frequência de postagem, informações demográficas, tópicos, unigramas e bigramas.
Nambisan et al. (2015)	Identificação de ruminação sobre sono e dor entre postagens de indivíduos deprimidos
Santos, Funabashi e Paraboni (2020)	Classificação binária de postagens possivelmente depressivas do Twitter, em português do Brasil, através de recursos linguístico-computacionais (LIWC, word embeddings), TF-IDF e algoritmos de AM tradicionais (regressão logística, MLP) e classificação e usuários
Yazdavar et al. (2017)	Classificação de sintomas em postagens do Twitter por meio de extração semi-supervisionada de tópicos com lista-semente.
Yadav et al. (2020)	Aprendizado multi-tarefa com ajuste fino do BERT, para classificar sintomas depressivos do PHQ-9 e linguagem figurada no Twitter
Mann, Paes e Matsushima (2020)	Classificação multimodal para análise de depressão no Instagram, usando aprendizado profundo, classificação de imagem e classificação textual com *features* diversas como *word embeddings*, TF-IDF e LIWC.

3 MATERIAIS E MÉTODOS

No decorrer da pesquisa notou-se que a **explicabilidade** seria uma característica altamente desejada para o classificador de PPD. Como o objetivo final do classificador automático gerado neste projeto é apontar indícios de que uma postagem é PPD, ser capaz de identificar as sentenças que levaram a esta classificação é muito importante para analisar a transparência e a confiabilidade do modelo gerado. Assim, buscou-se criar um classificador automático de postagens depressivas com boa capacidade explicativa, por meio de engenharia de *features*, que pudessem ser interpretadas sem grande esforço analítico, de forma a determinar possíveis fatores que levaram a uma dada classificação.

Tendo em vista a busca por explicabilidade, escolheu-se treinar modelos de aprendizado de máquina tradicionais baseados em engenharia de *features*, ao invés de técnicas de aprendizado profundo. Estudo anterior (MANN; PAES; MATSUSHIMA, 2020) aponta que modelos de aprendizado tradicionais são competitivos com modelos de aprendizado profundo para a tarefa de classificação de depressão, e são mais facilmente explicáveis.

O modelo proposto é, então, composto de 9 classificadores, cada um responsável pela classificação de um sintoma avaliado pelo PHQ-9[1], de forma a gerar uma classificação de postagem como PPD ou não. São os sintomas: (1) falta de interesse, (2) tristeza/humor depressivo, (3) desordem de sono, (4) falta de energia, (5) desordem alimentar, (6) baixa auto-estima, (7) problemas de concentração, (8) hiperatividade/baixa atividade e (9) pensamentos de suicídio. A Tabela 2 ilustra algumas possíveis formas de realização de cada um desses sintomas, inspiradas nas palavras e expressões presentes na lista semente gerada neste trabalho (veja Tabela 5).

Tabela 2 – Exemplos fictícios criados a partir da lista semente.

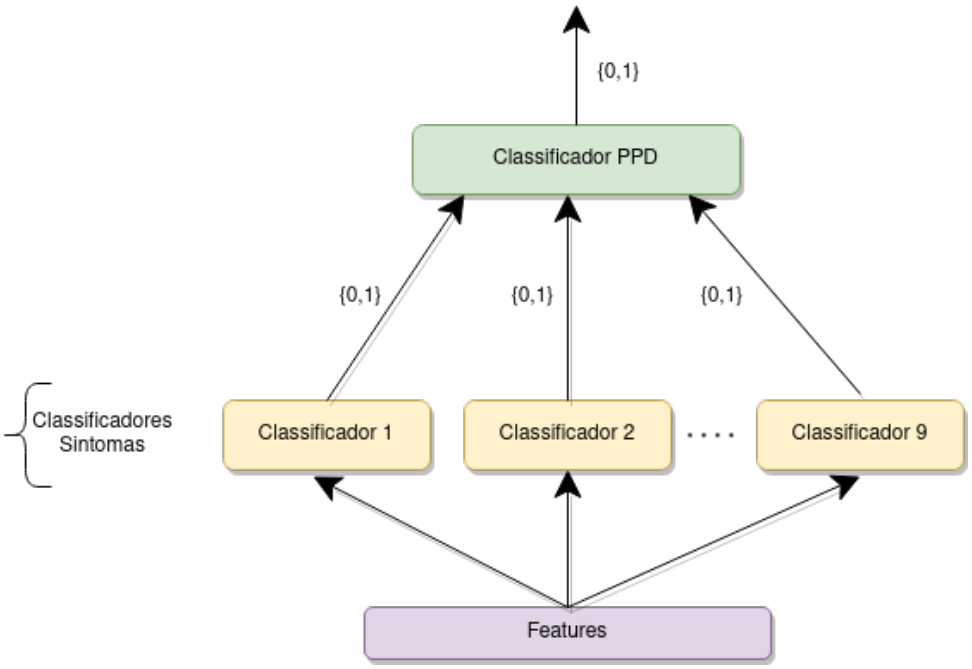
sintoma	**exemplo de texto**
falta de interesse	minha vida é um tédio
tristeza/humor depressivo	minha depressão voltou
desordem de sono	faz três dias que não durmo direito
falta de energia	não tenho vontade de levantar da cama
desordem alimentar	estou cada dia mais gordo
baixa auto-estima	eu me odeio
problemas de concentração	eu vou a aula, mas não consigo prestar atenção em nada
hiperatividade/baixa atividade	minha ansiedade voltou
pensamentos de suicídio	eu quero morrer

Fonte: autoria própria

[1] O *Patient Health Questionnaire* (PHQ-9) (<https://images.app.goo.gl/G6VoWfHL2Y4yVvLaA>) é um questionário de 9 itens usado para diagnóstico de depressão, voltado para rápida aplicação por clínicos gerais.

A partir desse conjunto de 9 sintomas, 9 classificadores foram treinados e utilizados em conjunto para determinar se uma dada postagem, ou um dado perfil de usuário, é ou não PPD. A Figura 5 ilustra o modelo proposto.

Figura 5 – Modelo proposto para a classificação de PPD com base na identificação dos 9 sintomas da PHQ-9.



Fonte: autoria própria

Esta estrutura visa proporcionar maior capacidade explicativa e alinhamento com práticas vigentes no campo de saúde mental, de forma que possa-se indicar a presença de um ou mais sintomas como fatores determinantes para a classificação da postagem e, consequentemente, do perfil como PPD.

Para o desenvolvimento e a avaliação destes modelos, 5 etapas foram realizadas: (1) rotulação fraca do *corpus* de (SANTOS; FUNABASHI; PARABONI, 2020), (2) pré-processamento do *corpus* para extração de *features*, (3) treinamento dos classificadores de sintomas, (4) treinamento dos classificadores de PPD por estratégia ensemble e (5) avaliação. As três primeiras etapas são descritas nas subseções a seguir e a avaliação dos modelos é apresentada no capítulo 4.

3.1 Rotulação fraca

Com o intuito de realizar a identificação automática de sintomas de depressão adotou-se a estratégia de rotulação fraca.[2] Tal estratégia foi adotada porque no momento do desenvolvi-

[2] A rotulação fraca é uma técnica para rotulação automática de um *corpus* a partir de casamento de padrão de uma lista de termos com os textos a serem rotulados.

mento desse projeto não havia, disponível livremente, um *corpus* anotado com sintomas da PHQ-9 para o português do Brasil.

A partir do *corpus* de (SANTOS; FUNABASHI; PARABONI, 2020), descrito na seção 3.1.1, uma rotulação fraca foi aplicada para anotar os sintomas. Neste processo de rotulação fraca, os novos rótulos (um dos sintomas da PHQ-9) de uma dada postagem são atribuídos com base em dois fatores: (i) a classe atribuída para a tarefa de detecção de depressão, e (ii) a presença de palavras-chave associadas a um dado sintoma. As palavras-chave, no caso, são as definidas na lista semente descrita na seção 3.1.2.

Assim, apenas as postagens rotuladas como positivas (PPD) no *corpus* de (SANTOS; FUNABASHI; PARABONI, 2020) foram processadas para associar, via rotulação fraca, a(s) classe(s) do(s) sintoma(s) que contém. Ao final desse processo, foram obtidas de 1.400 a 17.218 instâncias para os 9 sintomas da PHQ-9, conforme detalhado na Tabela 3.

Tabela 3 – Quantidade de instâncias para cada sintoma

Sintoma	**# instâncias**
1 - falta de interesse	5.052
2 - tristeza/humor depressivo	17.218
3 - desordem de sono	1.400
4 - falta de energia	4.854
5 - desordem alimentar	1.616
6 - baixa auto-estima	7.380
7 - problemas de concentração	2.162
8 - hiperatividade/baixa atividade	4.146
9 - pensamentos de suicídio	10.190

Assim, por exemplo, a postagem fictícia ilustrada na Tabela 4 seria rotulada como uma instância representativa tanto do sintoma "falta de interesse", devido à presença da palavra "tédio", como do sintoma "pensamentos de suicídio", devido à presença da expressão "quero morrer", assumindo que esta seja de autoria de um usuário da classe positiva no *corpus* original.

Tabela 4 – Exemplo fictício de uma postagem rotulada com os sintomas "falta de interesse" e "pensamentos de suicídio"

Minha vida é um tédio, eu quero morrer!

3.1.1 *Corpus* de (SANTOS; FUNABASHI; PARABONI, 2020)

Em (SANTOS; FUNABASHI; PARABONI, 2020), os autores coletaram um *corpus* do Twitter que consiste em usuários que relataram terem sidos diagnosticados com transtornos de saúde mental por profissionais da saúde, ou que relataram terem iniciado os tratamentos para uma dessas condições. Para tanto, a coleta se baseou em termos relacionados à saúde mental (como "depressão" e "ansiedade") e ao diagnóstico, tratamento, ou o uso de medicamentos antidepressivos.

Em seguida, todas as mensagens coletadas foram manualmente inspecionadas para filtrar aquelas que pareciam suficientemente genuínas. As 3.200 mensagens mais recentes dos autores foram, então, rotuladas apenas no nível do usuário. Para cada autor selecionado, as mensagens foram examinadas de modo a identificar o momento específico em que o diagnóstico ou tratamento foi iniciado, e para destacar o subconjunto de mensagens que foram publicadas antes desse evento.

Criou-se também um grupo de controle constituído por usuários que manifestaram interesse em questões de saúde mental como (i) uma preocupação geral (por exemplo, promovendo a campanha de prevenção ao suicídio "Setembro Amarelo"), (ii) uma preocupação em relação a uma pessoa particular que sofria de um problema de saúde mental (por exemplo, um amigo), ou (iii) por ser um estudante de psicologia com um interesse no tema da depressão. As mensagens obtidas em resposta a essas consultas foram inspecionadas manualmente para remover usuários diagnosticados ou tratados para problemas de saúde mental.

Por questões de sensibilidade e confidencialidade dos dados, não é possível apresentar neste documento nenhuma postagem do *corpus* de (SANTOS; FUNABASHI; PARABONI, 2020). Contudo, vale mencionar que a classificação de uma postagem estava associada ao usuário e não ao conteúdo da postagem. Dessa forma, mesmo as postagens sem conteúdo relativo a algum indício de sintoma está rotulada, nesse *corpus*, como sendo da classe positiva (é PPD) se o autor da postagem for um usuário rotulado como PPD. Essa característica pode significar um ruído para os modelos treinados no presente trabalho, que visa primordialmente a rotulação de sintomas presentes em postagens. Para tentar contorná-la, a rotulação das instâncias usadas no treinamento dos modelos considerou, também, a presença de palavras-chave associadas aos sintomas, como explicado a seguir.

3.1.2 Lista semente de (YAZDAVAR et al., 2017) traduzida para o português

Com a ajuda de um profissional de psicologia clínica, em (YAZDAVAR et al., 2017) foi produzida uma lista de termos relacionados a cada sintoma de depressão avaliado pelo PHQ-9. [3] Para tanto, os construtores da lista usaram a ferramenta Big Huge Thesaurus[4] e o dicionário colaborativo de gírias Urban Dictionary[5] para ajudar na busca por sinônimos. O léxico final contém mais de 1.620 entradas em inglês relacionadas à depressão. Além dos 9 sintomas da PHQ-9, os autores consideraram um décimo em seu léxico: medicamentos relativos a depressão.

A partir desta lista original (em inglês) foi gerada uma lista traduzida para o português. Dois nativos do português, com bons conhecimentos em inglês, traduziram as 1.620 entradas da lista original por meio de consultas ao tradutor automático Google Translate[6]. Nos casos em

[3] Disponível em: <https://github.com/yazdavar/Social-media-Depression-Detector/blob/master/depression_lexicon.json>
[4] Disponível em: <https://words.bighugelabs.com/>
[5] Disponível em: <https://www.urbandictionary.com/>
[6] Disponível em: <https://translate.google.com/>

que o tradutor não forneceu saída adequada a partir da entrada fora de contexto, buscou-se uma sentença completa na qual a entrada era usada em um contexto relevante ao sintoma[7].

Ao final deste processo, um conjunto de 1.213 *seeds*, em português, foi obtido. Diferenças no número de *seeds* em relação ao conjunto original têm duas causas. Primeiro, o conjunto continha expressões coloquiais (como "blothpick"), que nem sempre têm traduções óbvias para o português. Tais expressões coloquiais foram, então, desconsideradas. Outro fator para a diferença no número de *seeds* é que várias entradas diferentes no inglês levaram à mesma tradução. Além disso, boa parte dos termos do léxico original têm gênero neutro (como "*depressed*"), resultando na geração de mais termos correspondentes na tradução para o português ("deprimido" e "deprimida", por exemplo).

A Tabela 5 traz exemplos de palavras e expressões associados a cada um dos sintomas, que no léxico original e no traduzido são identificados como "sinais" (*signals*).

Tabela 5 – Exemplos de entradas no léxico de sintomas traduzido para o português

Sinal	Sintoma associado	Exemplo	n° de termos
1	falta de interesse	sem_graça, apatia, tédio	178
2	tristeza/humor depressivo	melancólico, deprimente, abatido	192
3	desordem de sono	insônia, sonolento, sem_dormir	82
4	falta de energia	não_tenho_força, preguiça, cansado	95
5	desordem alimentar	odeio_minhas_coxas, anorexia, barrigudo	146
6	baixa auto-estima	não_mereço, desprezível, eu_me_odeio	112
7	problemas de concentração	dispersa, distraído, desorientado	96
8	hiperatividade/baixa atividade	ansiedade, agitado, inquieto	81
9	pensamentos de suicídio	me_matar, merece_morrer, sono_eterno	104
10	medicamentos	lítio, alprazolam, citalopram	127

3.2 Pré-processamento e extração de *features*

Tratando-se de textos produzidos para mídias sociais, as postagens apresentam muitas abreviações, erros ortográficos, links e artefatos típicos de uma dada plataforma (por exemplo, menções no formato "@username" no Twitter). Estas características constituem ruído e afetam negativamente a extração de *features* e posterior treinamento de modelos.

Para tentar reduzir esses "ruídos", o *corpus* foi processado pela ferramenta de (BERTAGLIA; NUNES, 2016), o Enelvo[8], que faz: (1) normalização de abreviações (por exemplo, "vc" é normalizado para "você"), (2) correção de erros ortográficos e (3) identificação de ruídos (por exemplo, urls são todas normalizadas para a forma "url"), que foram subsequentemente removidos.

[7] Por exemplo, o tradutor automático gerou a equivalente "blues" para a palavra "bluesy", mas quando a mesma palavra foi empregada na sentença "I'm feeling bluesy" o tradutor foi capaz de gerar a tradução "Estou me sentindo triste".

[8] Disponível em: <vhttps://thalesbertaglia.com/enelvo/>

Após esse pré-processamento, as seguintes *features* usadas no treinamento do modelo foram extraídas:

LIWC – Utilizando o LIWC para o português[9], foi feita a contagem do número de termos, em cada postagem, que pertenciam às seguintes categorias: *family* (família), *anx* (ansiedade), *sad* (tristeza), *ingest* (alimentação), *work* (trabalho), *money* (dinheiro), *death* (morte), *friend* (amizade), *health* (saúde). Dessa forma, visou-se capturar informação a respeito de fatores sociais, econômicos e outros relacionados a sintomas somáticos como alimentação e saúde (termos neste tópico geralmente tratam da saúde física). Ao final deste processo, nove *features* binárias indicando presença (1) ou ausência (0) de termos em cada categoria foram geradas. A Tabela 6 exemplifica alguns termos de cada categoria.

Tabela 6 – Exemplos de termos para categorias do LIWC

Categoria	**Termos**
family	avó, casando-se, pais
anx	aflito, agitado, agoniada
sad	abandonado, arrependimento
ingest	assa, banha, bebendo
work	estudou, exige, desempregado
money	arruinada, conta, d espesas
death	morte, matar, suicida
friend	colega, familiarizado
health	adoentados, álcool

TF-IDF – Foi feita a extração da frequência de cada termo, em cada postagem, com contrapeso de acordo com a frequência de postagens como um todo (TF-IDF), com o intuito de capturar a temática e características de uma dada postagem que a diferencia das demais.

O valor de TF-IDF para uma dada palavra (termo) é o produto de dois fatores: (1) a frequência do termo em um dado documento de texto (*Term frequency*, TF), e (2) a frequência inversa da presença do termo em todos os documentos (*Inverse Document Frequency*, IDF). Para TF, temos:

$$TF(t,d) = \frac{freq(t,d)}{\sum_{t' \in d} freq(t',d)} \tag{3.1}$$

onde $freq(t,d)$ é o número de vezes que um dado termo t aparece em um documento d, e $\sum freq(t',d)$ é o somatório da frequência de todos os termos em um dado documento d.

Para IDF, dado um conjunto de documentos (*corpus*) D, temos:

$$IDF(t,D) = log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \tag{3.2}$$

[9] Disponível em: <http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc>

onde $|D|$ é o número total de documentos no *corpus*. Assim, IDF representa o inverso da frequência da presença de um dado termo t no conjunto de documentos, escalada logaritmicamente. Por fim, temos:

$$TFIDF(t, d, D) = TF(t, d) \times IDF(t, D) \tag{3.3}$$

A *feature* para um dado documento (no nosso caso, uma postagem) $d \in D$ consiste em um vetor numérico esparso contento os valores de TF-IDF para cada termo presente em D.

Polaridade (valência) – Outra *feature* utilizada neste trabalho foi a que indica a valência do sentimento da postagem como sendo negativa, neutra ou positiva. Neste trabalho, realizou-se o ajuste fino do modelo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) com o *corpus* TweetsentBR (BRUM; NUNES, 2018), tendo como resultado um classificador de polaridade que atingiu 67% de precisão no treinamento.

Vale notar que apesar da rede social (Twitter) usada para a construção do TweetSentBR ser a mesma do *corpus* usado neste projeto, o contexto do TweetSentBR é diferente e muito mais específico, uma vez que contém postagens relativas à programas de TV, visando capturar sentimentos da audiência. Assim, é preciso ter em mente que a forma como estes sentimentos são expressos nesse contexto específico podem não ser generalizáveis para tópicos de saúde mental.

O modelo treinado foi, então, usado para determinar a polaridade de uma postagem como a polaridade majoritária das sentenças que a compõem.[10] Uma *feature* categórica (-1 para negativa, 0 para neutra e 1 para positiva) foi gerada neste processo. Trata-se de uso de técnica de aprendizado profundo, o que potencialmente prejudicaria a explicabilidade do modelo, no entanto, tratando-se de classificação de polaridade, é relativamente simples analisar se a classe predita é congruente com o texto, descartando a necessidade de *features* com capacidade explicativa.

Embeddings – *Word Embeddings* são representações de palavras como pontos em um espaço vetorial n-dimensional, que é calculado de forma que palavras com significado similar estejam próximas. Desta forma, a *word embedding* de um dado termo é um vetor de dimensão n contendo as coordenadas da palavra no espaço. Utilizando as NILC-embeddings[11] (HARTMANN et al., 2017), modelo CBOW de dimensão 100, foram calculados os vetores para cada palavra em uma dada postagem. A representação da postagem foi gerada como a média dos vetores de palavra.

Além das *embeddings* pré-treinadas e de domínio geral do NILC, foram usadas também *embeddings* treinadas a partir do *corpus* de (SANTOS; FUNABASHI; PARABONI,

[10] Em caso de empate, adotou-se a polaridade neutra.
[11] Disponível em: <http://www.nilc.icmc.usp.br/embeddings>

2020) por meio do algoritmo Word2Vec (disponível na biblioteca `gensim`[12]), com peso determinado pelo valor TF-IDF de cada palavra aplicado ao vetor da mesma.

Assim, ao final deste processo, duas representações para a postagem foram usadas como *features*: uma baseada nas NILC-embeddings e outra baseada nas *embeddings* treinadas para o *corpus* de (SANTOS; FUNABASHI; PARABONI, 2020).

Etiquetas morfossintáticas – As etiquetas morfossintáticas são categorias associadas a uma palavra dada sua classe gramatical em um texto. Por exemplo, a palavra "estou" em "Eu estou em casa", tem as seguintes etiquetas associadas: verbo, 1ª pessoa, presente do indicativo, singular. Já a palavra "Eu" tem as seguintes etiquetas: pronome, 1ª pessoa, singular, e mf (tem mesma forma para gênero masculino e feminino).

Foram extraídas etiquetas morfossintáticas por meio de rotulador de *part-of-speech*. No caso, escolheu-se o rotulador da ferramenta Apertium[13], pois dentre os recursos para a língua portuguesa investigados, este ofereceu etiquetas mais detalhadas (por exemplo, foi a única ferramenta capaz de informar quando um dado verbo estava flexionado na primeira pessoa, em adição à etiqueta para verbo). No total, após a extração das etiquetas, foram identificadas 71 categorias morfossintáticas diferentes nos documentos que compõem o *corpus*. Assim, para cada postagem, 71 *features* numéricas indicam a frequência de cada etiqueta possível naquela postagem.

LDA – Por fim, a última *feature* considerada neste trabalho foi motivada pela análise de tópicos recorrentes em postagens de saúde mental. Utilizou-se a implementação da biblioteca `sklearn` do algoritmo *Latent Dirichlet Allocation* (LDA) para a extração de 10 tópicos no *corpus*, e subsequentemente foi feito o cálculo da chance de cada postagem pertencer aos tópicos, resultando em 10 *features* numéricas.

Assim, por exemplo, a partir do exemplo fictício da Tabela 4, algumas das *features* geradas seriam as ilustradas na Tabela 7.

Tabela 7 – Exemplo de *features* que poderiam ser extraídas para o exemplo fictício da Tabela 4

LIWC									
family	anx	sad	ingest	work	money	death	friend	health	Polaridade
0	0	0	0	0	0	1	0	0	-1

TFIDF (vetor esparso de tamanho 9050)
..., 0.38,...,0.26,...,0.36, ... , 0.56, ...,0.40, ... , 0.56, ..., 0.20

Etiquetas morfossintáticas (ilustrando apenas as que aparecem na postagem da Tabela 4)																	
det	pos	f	sg	n	vbser	pri	p3	ind	m	cm	prn	tn	p1	mf	vblex	inf	sent
2	1	2	7	2	1	2	1	1	2	1	1	1	2	1	2	1	1

[12] Disponível em: <https://radimrehurek.com/gensim/>
[13] Disponível em: <https://github.com/apertium>

Também experimentou-se com a redução da dimensionalidade do espaço de *features* utilizando *Principal Component Analysis* (PCA), com o intuito de melhorar a performance dos modelos gerados. Trata-se de possível caso de *trade-off* entre explicabilidade e performance, visto que o espaço de *features* reduzido, embora parcialmente explicável por meio de análise das combinações lineares dos componentes relevantes, é consideravelmente menos claro que o espaço original. O número de componentes usado foi escolhido com base no seguinte critério: os componentes devem explicar cumulativamente 90% da variância original, ou serem no máximo 50.

3.3 Treinamento dos modelos

Esta seção descreve brevemente os algoritmos de aprendizado de máquina clássicos usados nos experimentos para o treinamento dos classificadores de sintomas e classificador de PPDs.

- **Support Vector Machine**: Busca encontrar um hiperplano em um espaço n-dimensional de *features* que separe as amostras de acordo com suas classes da melhor maneira possível (maximizando a margem).
- **Regressão logística**: Análise de regressão que modela a probabilidade de um exemplo pertencer a uma dada classe, dada uma ou mais variáveis independentes. Para fins de classificação, a classe de maior probabilidade é a predita.
- **Multilayer Perceptron**: Rede neural *feed-forward* composta por ao menos 3 camadas de perceptrons (sendo uma delas para entrada e outra para saída). Dada uma função de ativação e pesos aleatórios, o modelo realiza iterações de *feed-forward* e subsequente *back propagation*, ajustando os pesos para a tarefa de classificação.
- **Random Forest**: Método *ensemble*, que consiste no treinamento de múltiplas árvores de decisão. No caso de tarefas de classificação, como a proposta, a classe predita pelo modelo é a mais predita pelos modelos que a compõem.

3.4 Classificador de PPD

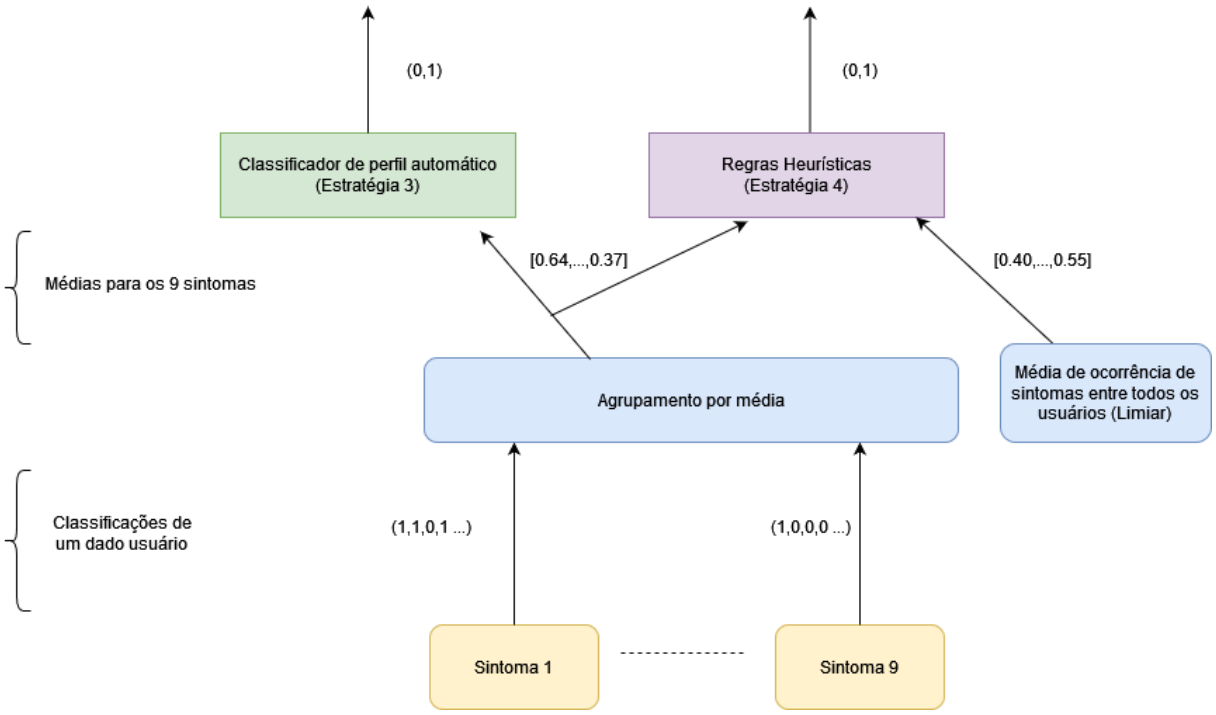
Foram investigadas 5 estratégias para a classificação de PPD, sendo que 3 destas aplicam classificação de postagens por meio do modelo proposto (utilizando os classificadores de sintomas) e as outras 2 extraem as *features* diretamente do *corpus* de (SANTOS; FUNABASHI; PARABONI, 2020), sem passar pelos classificadores intermediários.

- Estratégia 1 - classifica a **postagem** como PPD usando todas as *features* (sem passar pelos classificadores de sintomas).

- Estratégia 2 - classifica a **postagem** como PPD usando os classificadores de sintomas e um classificador automático (*ensemble*).
- Estratégia 3 - classifica o **perfil** como PPD usando os classificadores de sintomas e um classificador automático (*ensemble*).
- Estratégia 4 - classifica o **perfil** como PPD usando os classificadores de sintomas e regras heurísticas baseadas em definição médica de depressão.
- Estratégia 5 - classifica o **perfil** como PPD usando todas as *features* (sem passar pelos classificadores de sintomas).

Além da classificação das postagens como PPD ou não, este trabalho investigou também como os classificadores de sintomas poderiam ser usados para a classificação do perfil do usuário na RSO como PPD ou não. A Figura 6 ilustra as estratégias 3 e 4 propostas para esse fim.

Figura 6 – Estratégias propostas para a classificação em nível de usuário com base na identificação dos 9 sintomas da PHQ-9.



Fonte: autoria própria

4 RESULTADOS E DISCUSSÃO

Esta seção descreve os resultados dos experimentos realizados para: classificação de sintomas (seção 4.1) e classificação de PPD (seção 4.2).

4.1 Classificadores de sintomas

A Tabela 8 resume as configurações de melhor resultado, em termos de $F1$, para cada sintoma. Esses valores foram obtidos considerando-se a validação cruzada (5-*fold*) no treinamento dos modelos a partir do conjuntos de instâncias (Tabela 3) para cada sintoma, com diferentes combinações de *features*. A melhor configuração é apresentada como: Algoritmo (combinação de *features*).

Tabela 8 – Melhores resultados de $F1$ para cada sintoma

Sintoma	**Melhor valor de $F1$**	**Configuração do melhor resultado**
1 - falta de interesse	54,66%	LogReg (TF-IDF + LIWC + polaridade)
2 - tristeza/humor depressivo	56,09%	LogReg (TF-IDF + LIWC + polaridade)
3 - desordem de sono	56,47%	SVC (*embeddings* com peso)
4 - falta de energia	58,59%	LogReg (TF-IDF + LIWC)
5 - desordem alimentar	60,22%	LogReg (TF-IDF + LIWC + polaridade)
6 - baixa auto-estima	56,18%	SVC (todas as *features*)
7 - problemas de concentração	55,83%	LogReg (TF-IDF + LIWC + polaridade)
8 - hiperatividade/baixa atividade	58,43%	MLP (todas as *features*)
9 - pensamentos de suicídio	57,66%	LogReg (TF-IDF + LIWC)
Média	57,12%	–

A regressão logística (LogReg) foi o algoritmo de melhor desempenho na classificação de 6 das 9 classes de sintomas. As *features* TF-IDF e LIWC, combinadas ou não com polaridade, também se destacaram. A comparação direta com os trabalhos relacionados não é indicada neste caso, uma vez que idioma e modo de anotação (rotulação fraca) adotados aqui diferem dos trabalhos da literatura. Contudo, vale mencionar que este trabalho estende (SANTOS; FUNABASHI; PARABONI, 2020) e corrobora sua constatação de que regressão logística com TF-IDF tem desempenho melhor na tarefa de classificação de PPDs.

As *embeddings* geradas só apresentaram melhor performance em 1 dos sintomas, configurando *trade-off* entre performance e explicabilidade. Em todos os casos, as *embeddings* treinadas com o *corpus* de (SANTOS; FUNABASHI; PARABONI, 2020) apresentaram performance competitiva com os demais modelos, enquanto as pré-treinadas não atingiram performance melhor que chance aleatória. Uma possível explicação é a diferença de domínio uma vez as *embeddings* do NILC foram treinadas com textos extraídos da Wikipedia, Google News, repositórios acadêmicos, *ebooks*, etc., que apresentam linguagem e estrutura radicalmente

diferentes de postagens publicadas em redes sociais e coletadas visando a temática específica de saúde mental.

Modelos que usavam as demais *features* (etiquetas morfossintáticas e distribuição de tópicos) só superaram performance dos modelos que não as usavam em dois sintomas: "baixa auto-estima" e "hiperatividade / baixa atividade".

É importante ressaltar que a estratégia de rotulação fraca apresenta limitações. Apesar de ela se basear na presença de diagnóstico prévio e na ocorrência de palavras e expressões associadas aos sintomas, não é possível garantir que a postagem em questão relata mesmo o sintoma. A rotulação fraca é especialmente prejudicada pela presença de ambiguidade e pela subjetividade da tarefa, uma vez que nem toda postagem produzida por uma pessoa deprimida necessariamente evidencia sintoma de depressão.

Foi feita análise das classificações obtidas em busca de padrões. No caso do sintoma 3 "desordem do sono", por exemplo, observou-se que o classificador considera a palavra "cansado(a)" como forte indício de presença do sintoma, classificando muitas das postagens que incluem a palavra como indicativas do sintoma. Outros casos incluem "triste" para o sintoma 2 "tristeza/humor depressivo", e "açúcar" no caso de "desordem alimentar". Muitas das vezes, como exemplificado, estes termos estavam inclusos no léxico traduzido. Visto que, devido a estratégia de rotulação fraca, todos os textos do corpus contêm ao menos um termo contido no léxico, o fato dos classificadores identificarem um subconjunto destes como indicativos dos sintomas é relevante.

4.2 Classificadores de PPD

Como discutido na seção 3.4, o modelo proposto para a classificação de uma postagem como PPD ou não recebe as previsões dos classificadores de sintomas como entrada, caracterizando estratégia *ensemble*. Foram investigados modelos de regressão logística, SVM, MLP e Random Forest. Para fins de comparação, tomou-se como *baseline* os melhores resultados obtidos por (SANTOS; FUNABASHI; PARABONI, 2020) para cada tarefa de classificação da postagem ou do perfil do usuário na RSO.

4.2.1 Classificação de postagens

No caso da classificação de postagens individuais, a estratégia 2 (veja seção 3.4), quando aplicada a todo o *corpus*, gerou modelos triviais que classificavam quase todos os textos como pertencentes à classe majoritária (não PPD). Dessa forma, realizou-se nova rodada de experimentos com *undersampling* da classe majoritária de forma que os modelos fossem treinados com um *corpus* balanceado. O *corpus* original contém 603.965 postagens, sendo que 274.693 pertencem à classe positiva (PPD) e 317.114 à classe negativa. Após *undersampling*, o *corpus* totalizou 549.386 postagens, divididas igualmente entre as classes.

Os modelos resultantes desta segunda rodada de experimentos apresentaram performance semelhante a chance aleatória. Rodadas subsequentes de experimentos foram conduzidas, selecionando as 3, 2 e 1 melhores *features* segundo o teste chi2, sendo estes, em ordem crescente de relevância: sintoma 1, sintoma 4 e sintoma 6. Estas rodadas apresentaram performance similar às anteriores, o que é compatível com a observação de que nenhuma das *features* apresentou valor-p menor que 0.05, de forma que não foi possível rejeitar a hipótese nula de que não há correlação entre *feature* e classe. A Tabela 9 compara a performance do classificador treinado por (SANTOS; FUNABASHI; PARABONI, 2020) com o gerado pela estratégia 2.

Tabela 9 – Comparação entre modelo proposto para classificação de postagens utilizando as classificações de sintomas e o *baseline* de (SANTOS; FUNABASHI; PARABONI, 2020)

Modelos	F1	Revocação	Precisão
Santos, Funabashi e Paraboni (2020)	69%	69%	69%
Estratégia 2 (Random Forest)	51%	50%	53%

A partir da análise desses resultados, sugere-se duas possíveis explicações para a baixa performance da estratégia 2 (baseada na classificação prévia dos 9 sintomas) na tarefa de classificação de postagens:

1. Esta estratégia apresentou performance pior para a classificação de postagens individuais pois as postagens carecem de contexto. Como observado na definição de depressão maior, esta é caracterizada pela manifestação de sintomas ao longo de um determinado período de tempo (no caso, 2 semanas), o que dificilmente pode ser capturado em uma única postagem.

2. O impacto do ruído gerado pelos classificadores de sintomas pode ter influenciado negativamente a classificação de postagens individuais.

4.2.2 Classificação de perfis de usuários na RSO

No caso de classificação de perfil do usuário, foram considerados todos os rótulos de todas as postagens de um determinado usuário, totalizando 9 *features* (uma para cada sintoma). Visto que a quantidade de postagens pode variar muito entre os usuários, a presença (valor da *feature* igual a 1) ou não (valor da *feature* igual a 0) de um sintoma foi determinada com base em um limiar. Assim, determinou-se como critério de presença de sintoma um conjunto de valores que serviriam como limiar, ou seja, se o valor de um sintoma for menor que o valor de limiar adotado para o mesmo, considera-se que o sintoma não foi detectado no usuário.

Inicialmente, considerou-se adotar como valor de limiar o valor médio de um dado sintoma entre todos os usuários. No entanto, dados os valores de F1 para os classificadores de sintomas, espera-se ruído nas classificações geradas, e ajustes no valor de limiar pode

compensar viés dos classificadores para uma dada classe (ao exigir valor médio maior/menor para determinar presença). Desse modo, tomando-se como base o valor médio M para cada sintoma, investigou-se limiares no intervalo [-0,15+M, 0,15+M], em incrementos de 0,01. No geral, o único limiar que produziu melhor resultado foi M (a média).

No caso da estratégia 4, procurou-se alinhar a heurística adotada com a definição de depressão maior usada no PHQ-9: presença de 5 sintomas, sendo ao menos um destes tristeza/humor deprimido ou perda de interesse, em um período de 2 semanas. Dada a natureza do *corpus*, não foi possível incorporar o aspecto temporal da definição pois, apesar das postagens de cada usuário estarem em ordem de publicação, carecem de data, não sendo possível dividi-las em intervalos de 2 semanas.

A Tabela 10 compara os melhores modelos das estratégias 3 e 4 com os resultados obtidos por (SANTOS; FUNABASHI; PARABONI, 2020).

Tabela 10 – Comparação entre os modelos propostos para classificação de usuários utilizando as classificações de sintomas e o *baseline* de (SANTOS; FUNABASHI; PARABONI, 2020)

Modelos	F1
Santos, Funabashi e Paraboni (2020)	64%
Estratégia 3 (LogReg)	64%
Estratégia 3 (Random Forest)	59%
Estratégia 4 (Cutoff na média)	48%

A estratégia 3 gerou modelos comparáveis aos resultados obtidos por (SANTOS; FUNABASHI; PARABONI, 2020), atingido F1 de 64% com Regressão Logística e 59% com Random Forest, o que sugere que, apesar do método baseado em heurística (estratégia 4) não produzir resultados satisfatórios, as classificações de sintomas contêm informação relevante para a tarefa de classificação de perfil de usuário. Modelos gerados por MLP e SVC, no entanto, produziram modelos triviais. Acredita-se que o impacto do ruído gerado pelos classificadores de sintomas pode ter sido diluído no grande volume de postagens usadas para a classificação de cada usuário (média de 2708 postagens por usuário).

4.3 Resultados dos demais modelos

Os modelos que não utilizaram os classificadores de sintomas (estratégias 1 e 5) apresentaram baixa performance, próximo de chance aleatória para ambas as tarefas (classificação de perfis de usuários e de postagens). No caso do classificador em nível de usuário, os textos das postagens foram concatenados, e foi calculada a média das frequências no caso das etiquetas morfossintáticas. A partir destes *inputs*, o modelo foi treinado usando os mesmos conjuntos de *features* usados para a classificação de sintomas. As tabelas 11 e 12 trazem comparações entre

os modelos tradicionais gerados e o melhores modelos de (SANTOS; FUNABASHI; PARABONI, 2020) para as respectivas tarefas.

Tabela 11 – Comparação entre modelo tradicional e *baseline* para classificação de postagens

Modelos	F1	Revocação	Precisão
Santos, Funabashi e Paraboni (2020)	69%	69%	69%
LogReg (todas as *features* - estratégia 1)	47%	51%	53%

Tabela 12 – Comparação entre modelo tradicional e *baseline* para classificação de perfis de usuários

Modelos	F1
Santos, Funabashi e Paraboni (2020)	64%
LogReg (todas as *features* - estratégia 5)	36%

Uma possível razão para a piora considerável de performance é o maior conjunto de *features*. O maior volume de textos usado para estas tarefas, em comparação aos *corpora* usados para o treinamento de classificadores de sintomas, leva à geração de um espaço de *features* significativamente maior. Nestas condições, a redução de dimensionalidade por PCA, dadas as restrições impostas, pode não ter sido capaz de capturar informação suficiente.

4.4 Análise de explicabilidade

Entre os objetivos desta pesquisa, buscou-se elaborar um conjunto de *features* com potencial explicativo. Em alguns casos isto é evidente, como a presença de categorias do LIWC como "money" e "work", relativos a fatores sociais e econômicos, ou a polaridade de uma sentença. Contudo, para outras *features* como as etiquetas morfossintáticas e os tópicos, o significado das mesmas não é óbvio, necessitando interpretação. Esta interpretação poderia ser estabelecida de antemão (ou seja, seria formado consenso entre os usuários do modelo, após análise qualitativa), ou poderia ser realizada de forma *ad-hoc* por profissional de saúde mental, dado o contexto de cada paciente.

Visto a ênfase colocada na explicabilidade dos modelos gerados durante a pesquisa, julgou-se inclusa no escopo da mesma a realização de análise qualitativa, de forma a avaliar se as *features* propostas de fato contém informação suficiente para a explicação das classificações. Como foi aplicada técnica de redução de dimensionalidade por PCA, foi necessária análise adicional dos componentes principais. Dada a natureza sensível dos dados, esta seção não apresenta exemplos de postagens, mas traz as conclusões generalizadas a respeito da análise.

Como esperado, a análise dos 10 termos com maiores valores de TF-IDF em cada postagem demonstra que estes muitas vezes contém pistas sobre o assunto da mesma, e ocasionalmente incluem termos relativos aos sintomas, como "sono" e "chorando" , mesmo com presença elevada destes termos nos documentos do *corpus*, o que acarreta maior penalidade

IDF. Notou-se também que termos indicativos de gênero ou na primeira pessoa costumam apresentar maiores valores de TF-IDF.

A análise do primeiro componente dos classificadores indica que dos 10 principais fatores, 8 são categorias do LIWC, sendo responsáveis por grande parte da variância. Isto reforça a importância destas *features* para a tarefa de classificação. Em retrospecto, estas *features* não deveriam ter sido inclusas no processo de redução de dimensionalidade, visto que têm pouco impacto na alta dimensão do espaço de *features* original, e apresentam alta capacidade explicativa.

Após a avaliação dos 10 principais termos dos tópicos gerados para todas as postagens rotuladas para cada sintoma, concluiu-se que eles são adequados para fins explicativos, mesmo com a repetição de alguns termos. Por exemplo, todos os tópicos relativos ao sintoma 3 (desordem de sono) têm, entre seus principais termos, palavras relacionadas a "insônia", como ilustrado na Tabela 13[1]. O tópico 1 contém termos como "cochilo" e "tarde", sugerindo sono ao longo do dia, enquanto os tópicos 8 e 9 contêm os termos "preciso", "consigo" e "quero", juntamente com "sono" e "dormir", indicando incapacidade de dormir quando desejado. Essa co-ocorrência de termos nos tópicos indica também que o uso de n-gramas na construção do léxico de (YAZDAVAR et al., 2017) é de fato interessante para a identificação de sintomas.

Tabela 13 – Tópicos gerados para o sintoma 3 ("desordem de sono"), com termos que indicam significado interpretável

Tópico	Termos
1	pra, dormir, acordada, vou, ficar, tirar, **tarde**, **cochilo**, agora, porque
2	url, insônia, agora, dormir, amor, vida, ter, pessoas, nome, noturna
3	**noturno**, **modo**, acordado, ter, **twitter**, coisa, username, motivo, to, **cama**
4	username, acordada, noite, ainda, bem, durma, acordado, pra, boa, dia
5	**acordada**, to, **acordado**, fazendo, pra, **hora**, url, username, porque, **horas**
6	dormir, acordada, dia, to, insônia, acordado, **kar****, todo, sono, bom
7	pra, username, jazz, insônia, acordada, vou, hoje, **car****, acordado, tudo
8	dormir, **consigo**, agora, **preciso**, pra, noite, **quero**, acordado, ficar, insônia
9	dormir, **quero**, hoje, tarde, **consigo**, pra, **sono**, acordar, dias, porque
10	insônia, pra, dia, username, acordado, url, porque, vai, agora, kkk

4.5 Outros resultados

Durante o desenvolvimento da pesquisa, *corpora* adicionais foram coletados, mas estes foram reservados para uso futuro pois não foi possível finalizar a tarefa de anotação desses *corpora* dentro da vigência do trabalho. Tais *corpora* são compostos de postagens coletadas das seguintes redes sociais:

[1] Foi feita remoção de *stopwords* (palavras muito comuns na linguagem), no entanto alguma expressões como "pra"e "porque"não estavam inclusas na lista usada, portanto constam nas tabelas.

Tabela 14 – Tópicos gerados para o sintoma 1 (“falta de interesse”), com termos que indicam significado interpretável

Tópico	Termos
1	tédio, chato, vai, car****, pra, ódio, c*, porque, ser, quero
2	ódio, **exausta**, **cansaço**, pessoas, todo, ser, dia, **hoje**, ter, alguém
3	cansada, to, cansado, ódio, vou, porque, tudo, dormir, fazer, pra
4	tão, cansada, chato, ai, pra, to, **desanimada**, vida, fazer, tudo
5	ódio, pu**, chato, pariu, pra, ser, mano, ainda, porque, cara
6	cansada, to, dia, vou, chato, semana, cansado, ódio, hoje, pra
7	pra, chato, nada, **mentalmente**, chorar, ódio, tudo, exausto, porque, dia
8	ódio, to, amor, pra, hoje, chato, força, kkk, tanto, sim
9	graça, tudo, gente, ódio, chato, dia, pessoa, toda, **cheio**, **saco**
10	pra, cansada, ódio, porque, cansado, chato, pessoa, ser, mim, to

Tabela 15 – Tópicos gerados para o sintoma 2 (“tristeza/humor deprimido”), com termos que indicam significado interpretável

Tópico	Termos
1	triste, ai, dia, porque, cara, bem, nada, pra, ser, to
2	pra, ai, triste, hoje, dia, vai, ter, ser, ver, gente
3	ai, triste, pra, porque, vou, to, kkk, dia, hoje, vai
4	triste, tristeza, to, **desespero**, vai, **medo**, tão, vou, ai, tudo
5	ai, porque, triste, pra, chorando, dor, amor, horrível, coração, deus
6	horrível, triste, ai, ser, fico, depressão, pessoas, porque, sei, ainda
7	chorando, ai, to, kkk, puta, casa, pariu, amo, triste, agora
8	ai, triste, pra, fazer, porque, tudo, kkk, vai, chorando, ser
9	ai, gente, pra, **agonia**, depressão, **ninguém**, fiquei, desgosto, ainda, fica
10	pra, triste, to, ai, pessoa, gente, depressão, bem, ser, tudo

Tabela 16 – Tópicos gerados para o sintoma 4 (“falta de energia”), com termos que indicam significado interpretável

Tópico	Termos
1	cansada, to, pra, preguiça, tão, tudo, cansaço, dormir, cansado, ai
2	mulher, **lerdo**, tudo, gente, homem, vagabundo, ah, falsa, kkk, vagabunda
3	preguiça, to, pra, cansado, dia, **hoje**, cansada, ir, **nada**, casa
4	exausta, pra, falso, bem, kkk, porque, gente, preguiça, **lerda**, falsa
5	preguiça, pra, fazer, to, amiga, vou, porque, mãe, vagabundo, cabelo
6	vida, preguiça, nunca, ser, pra, cansativa, porque, dias, dia, tão
7	fraco, ser, pra, vagabunda, ponto, vai, **forte**, **frágil**, mundo, falsa
8	pra, cansada, fraca, falsa, ser, **lento**, tão, sim, ter, vagabundo
9	vou, preguiça, **lenta**, pu**, pariu, hoje, falso, pra, tédio, fazer
10	tédio, preguiça, pra, faz, vai, fazer, to, cara, alguém, ficar

Tabela 17 – Tópicos gerados para o sintoma 5 (“desordem alimentar”), com termos que indicam significado interpretável

Tópico	Termos
1	quilogramas, frágil, masculinidade, pra, coisas, engordar, quer, fazem, **feio**, peso
2	pra, **emagrecer**, to, **5kg**, perder, agora, pesado, **1kg**, frágil, ano
3	**emagrecer**, pra, vou, **preciso**, to, cara, frágil, educação, porque, faz
4	ter, pra, cocaína, quilos, to, queria, fazer, bunda, 9kg, estar
5	ser, vai, pra, pesado, magra, magro, nunca, tênue, linha, corpo
6	mulher, tudo, ideai, pra, frágil, deixa, fácil, ser, fazer, gordo
7	pra, gordo, porque, pesado, vou, bem, emagrecer, vai, **comer**, dia
8	pesado, fazer, kkk, engordar, pra, cara, tava, tanto, porque, acho
9	pesado, pornô, demais, kkk, tão, hoje, atos, en, adocem, pra
10	ser, gordo, magra, pesado, gente, porque, kkk, pra, hoje, bem

Tabela 18 – Tópicos gerados para o sintoma 6 (“baixa auto-estima”), com termos que indicam significado interpretável

Tópico	Termos
1	pra, pior, porque, horrível, ninguém, pessoa, vergonha, mano, mundo, gente
2	pior, ser, pra, kkk, vergonha, feio, ainda, horrível, coisa, ter
3	pior, pra, vergonha, vida, cara, ser, vai, fazer, coisa, ano
4	horrível, pior, **feio**, ter, fracasso, **foto**, ser, vai, tudo, gente
5	vergonha, feio, pior, gente, **desistir**, pessoa, pra, porque, ser, ficar
6	pior, pra, alguém, ter, verdade, to, horrível, pessoa, gente, porque
7	pra, feio, pior, dia, vergonha, pessoas, **desistir**, ter, tudo, vou
8	pior, feio, ser, pu**, horrível, kkk, pariu, vergonha, vai, ter
9	foda, se, pior, **desistir**, pra, vergonha, tudo, vou, pessoa, vai
10	**vergonha**, passar, pior, cara, kkk, pra, **alheia**, porque, ai, tão

Tabela 19 – Tópicos gerados para o sintoma 7 (“problemas de concentração”), com termos que indicam significado interpretável

Tópico	Termos
1	ser, preocupado, vida, gente, **esquecimento**, afastar, amor, noção, pra, pessoa
2	preocupado, ser, kkk, **confusa**, **fiquei**, pra, perdido, **confuso**, caso, tão
3	grosso, afastar, noção, pra, vou, vai, pessoas, bem, preocupar, kkk
4	**esquecido**, pra, afastar, preocupado, perdido, bom, tudo, bem, tava, ter
5	afastar, confusa, pessoa, porque, ser, gente, ter, melhor, to, pra
6	to, pra, confusa, perdido, preocupado, tão, tudo, sei, quero, preocupar
7	pra, perdido, esquecido, afastar, kkk, vida, hoje, tanto, nunca, confusa
8	pra, perdido, porque, preocupar, grosso, dia, pessoas, ter, alguém, ser
9	pra, perdido, preocupar, noção, porque, vou, ter, mundo, vida, gente
10	noção, pra, gente, preocupado, perdido, vida, porque, vai, aqui, **sempre**

Tabela 20 – Tópicos gerados para o sintoma 8 (“hiperativade/baixa atividade”), com termos que indicam significado interpretável

Tópico	Termos
1	ansiedade, vai, pequeno, todo, **pânico**, hoje, pode, mundo, pra, depressão
2	**ansiedade**, pra, dormir, chato, vou, **crise**, porque, coisas, pessoa, pequeno
3	chato, pra, porque, to, ser, ansiedade, hoje, gente, fico, **estressado**
4	chato, pra, pequeno, ser, car****, tão, porque, vai, kkk, gente
5	ansiedade, pra, nervoso, tava, tremendo, chato, to, porque, gente, aí
6	ansiedade, pra, chato, tudo, porque, fazer, ter, dia, ser, **crises**
7	ansiedade, pra, gente, kkk, chato, vai, ser, ter, dia, pequeno
8	ansiedade, pra, dia, chato, atrasado, vou, ainda, kkk, porque, remédio
9	pra, **nervoso**, to, **rindo**, ansiedade, ansioso, tremendo, kkk, vai, pequeno
10	ansiedade, chato, crise, pu**, pra, depressão, pariu, vai, porque, ter

Tabela 21 – Tópicos gerados para o sintoma 9 (“pensamentos de suicídio”), com termos que indicam significado interpretável

Tópico	Termos
1	**morrendo**, to, pra, acabou, sono, kkk, saudade, **morrer**, aqui, agora
2	dor, cabeça, pra, to, porque, dia, **inferno**, hoje, muita, **cortar**
3	**inferno**, **acabou**, vai, mundo, todo, drama, porque, pro, pra, dia
4	**acabou**, pra, **inferno**, **morrer**, deus, hoje, **morto**, gente, cara, dia
5	dor, **morrendo**, sim, **morte**, porque, vai, **inferno**, **morrer**, ser, pra
6	dor, coração, ser, pra, **morte**, doeu, vai, ter, **morto**, vem
7	**morte**, **inferno**, **acabou**, pessoas, porque, bem, gente, anos, ainda, bolsonaro
8	**morrer**, vou, pra, quero, **matar**, vai, **cortar**, **inferno**, cabelo, **acabou**
9	**morrer**, vai, pra, tudo, ia, dor, **morte**, ser, tava, porque
10	**morte**, **morto**, **acabou**, anos, bandido, pra, hoje, porque, ex, **vida**

- **Reddit:** A plataforma Reddit [2] possui características que justificam sua inclusão, apesar da adoção significativamente menor por parte do público brasileiro. Em primeiro lugar, seu carácter anônimo facilita a discussão de tópicos sensíveis, como é o caso de temas associados a depressão. Além disso, a plataforma é composta por diversas comunidades bem definidas, o que permite rotulação automática. No caso, foram coletadas postagens da comunidade `r/desabafos` que tinham a *tag* “Depressão” associada às mesmas. Como grupo de controle, foram coletadas postagens da comunidade de uso geral `r/brasil`, que devem ser subsequentemente validadas de forma a garantir que não apresentam postagens que relatam sinais de depressão. Estas postagens foram coletadas por meio da API `pushift`[3]. O *corpus* contém 8.412 postagens, 76.867 sentenças e 1.681.495 *tokens*, datadas de 02 de fevereiro de 2020 até 27 de julho de 2021.

- **Facebook:** Postagens do Facebook foram coletadas por meio da ferramenta Crowdtangle[4].

[2] <https://reddit.com>
[3] Disponível em: <https://github.com/pushshift/api>
[4] Disponível em: <https://www.crowdtangle.com/>

Inicialmente, buscou-se por palavras-chave relacionadas a depressão (termos de busca final usados: "depressão", "suicídio", "me corto", "vontade de viver", "me matar", "quero morrer"). No entanto, ao contrário das demais redes sociais investigadas, a análise inicial sugeriu que os resultados retornados não foram satisfatórios, visto que não foi possível identificar PPDs dentre as mensagens coletadas.

Uma possível explicação é a natureza da plataforma como um espaço cultivado pelo usuário, de forma que expressões públicas de sintomas de depressão sejam incomuns. A adesão massiva do público brasileiro a esta plataforma motivou a busca por outra forma de coleta. No caso, foram feitas buscas similares, mas em páginas de segredos universitários, que fornecem um espaço anônimo, facilitando a expressão de sintomas de depressão assim como visto nas comunidades do Reddit. Estas páginas foram encontradas por meio de lista de universidades brasileiras[5], com a adição do termo "segredos" (por exemplo, "UFBA" + "segredos").

Apesar do ambiente anônimo, as postagens foram verificadas manualmente em busca de termos possivelmente identificadores, como nomes de universidades, departamentos, cursos, locais, etc. Estes termos foram, então, substituídos por *tags* genéricas (por exemplo, menções à UFSCar foram substituídas por tag <universidade>). Durante o processo de anonimização manual, foram anotados os termos substituídos, muitos deles comuns a várias postagens (por exemplo, menções a um mesmo curso), e foi criado um *script* para anonimização automática, caso deseje-se coletar mais postagens destas páginas no futuro.

As postagens coletadas foram feitas entre 31/12/2011 e 24/11/2020. O *corpus* final contém 705 postagens com cerca de 7.400 sentenças e 117.414 *tokens*,e está, neste momento, sendo anotado por uma equipe de especialistas em saúde mental vinculada ao projeto Amive.

[5] Disponível em: <https://pt.wikipedia.org/wiki/Lista_de_institui%C3%A7%C3%B5es_de_ensino_superior_do_Brasil>

5 CONCLUSÕES

Neste Trabalho de Conclusão de Curso foram avaliadas diversas estratégias de aplicação de PLN para classificação automática de PPD em RSO em português do Brasil, muitas das quais, até onde se pôde apurar nos trabalhos da literatura, não haviam sido investigadas até então. Estas incluem lacunas na investigação de *features* tradicionais (etiquetas morfossintáticas e tópicos), técnicas de rotulação automática (rotulação fraca) e a composição de modelos *emsemble* informados pela literatura médica, utilizando classificadores de sintomas.

Apesar de modelos para a classificação de PPDs em postagens individuais atingirem performance significativamente menor que o observado na literatura, os classificadores de sintomas e o classificador automático de perfis de usuários demonstram, respectivamente, a possibilidade de identificação automática de sintomas em postagens individuais e do uso destes sintomas para a classificação de perfis de usuários possivelmente depressivos na lingua portuguesa. Uma possível direção para trabalhos futuros seria a investigação destas estratégias em outras RSO e com outras estratégias de anotação de *corpus*, visto que a rotulação fraca apresenta limitações em termos de ruído, e o Twitter é uma RSO caracterizada por textos curtos.

Além disso, classificadores baseados em engenharia de *features* tradicionais podem ser desenvolvidos tendo sua explicabilidade em mente, característica essencial quando se considera a possibilidade de acionamento de intervenções de saúde mental em caso de classificação positiva. A análise qualitativa conduzida demonstra que boa parte do conjunto de *features* usado apresenta capacidade explicatória, embora observou-se *trade-offs* entre performance e explicabilidade dos modelos, especialmente se tratando da aplicação de técnica de redução de dimensionalidade.

Em relação aos experimentos com *corpus* de (SANTOS; FUNABASHI; PARABONI, 2020), algumas das conclusões do estudo original foram corroboradas, como a boa performance de modelos de regressão logística e da *feature* TF-IDF para a tarefa. Tendo em mente a escassez de recursos em língua portuguesa para a tarefa proposta, o compartilhamento deste conjunto de textos foi essencial para a condução dos experimentos. Isso demonstra a importância de construção de recursos compartilhados para a exploração de aplicações de PLN na saúde mental em português, aos quais este trabalho oferece contribuição por meio de disponibilização da lista de termos traduzidos de (YAZDAVAR et al., 2017) no github do LALIC: <https://github.com/LALIC-UFSCar/Amive-PLN/>. Como evidenciado pela análise dos tópicos gerados, que continham termos relevantes não inclusos no léxico traduzido, este recurso pode ser expandido e refinado em trabalhos futuros com auxílio de especialistas em saúde mental, assim como feito pelos autores do estudo original.

Vale mencionar que os demais *corpora* coletados neste projeto serão usados no escopo do projeto Amive para continuação e extensão da investigação da identificação de PPD em RSO usando recursos de PLN e AM iniciada neste trabalho.

REFERÊNCIAS

American Psychiatric Association et al. DSM-5: Manual diagnóstico e estatístico de transtornos mentais. [S.l.]: Artmed Editora, 2014. Citado 2 vezes nas páginas 11 e 17.

BERTAGLIA, T. F. C.; NUNES, M. d. G. V. Exploring word embeddings for unsupervised textual user-generated content normalization. In: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). [S.l.: s.n.], 2016. p. 112–120. Citado na página 27.

BRUM, H.; NUNES, M. d. G. V. Building a Sentiment Corpus of Tweets in Brazilian Portuguese. In: CHAIR), N. C. C. et al. (Ed.). Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018. ISBN 979-10-95546-00-9. Citado na página 29.

CHOUDHURY, M. D.; COUNTS, S.; HORVITZ, E. Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference. New York, NY, USA: Association for Computing Machinery, 2013. (WebSci '13), p. 47–56. ISBN 9781450318891. Disponível em: <https://doi.org/10.1145/2464464.2464480>. Citado 4 vezes nas páginas 12, 16, 17 e 21.

COPPERSMITH, G. et al. Natural language processing of social media as screening for suicide risk. Biomedical informatics insights, SAGE Publications Sage UK: London, England, v. 10, p. 1178222618792860, 2018. Citado na página 11.

EICHSTAEDT, J. C. et al. Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences, National Acad Sciences, v. 115, n. 44, p. 11203–11208, 2018. Citado 3 vezes nas páginas 16, 18 e 21.

HARTMANN, N. et al. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. 2017. Citado na página 29.

JI, S. et al. Supervised learning for suicidal ideation detection in online user content. Complexity, v. 2018, p. 1–10, 09 2018. Citado 3 vezes nas páginas 11, 15 e 21.

KROENKE, K.; SPITZER, R. L.; WILLIAMS, J. B. The phq-9: validity of a brief depression severity measure. Journal of general internal medicine, Wiley Online Library, v. 16, n. 9, p. 606–613, 2001. Citado na página 11.

MANN, P.; PAES, A.; MATSUSHIMA, E. H. See and read: Detecting depression symptoms in higher education students using multimodal social media data. In: Proceedings of the International AAAI Conference on Web and Social Media. [S.l.: s.n.], 2020. v. 14, p. 440–451. Citado 4 vezes nas páginas 12, 19, 21 e 23.

NAMBISAN, P. et al. Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter. In: IEEE. 2015 48th Hawaii International Conference on System Sciences. [S.l.], 2015. p. 2906–2913. Citado 2 vezes nas páginas 17 e 21.

SANTOS, W.; FUNABASHI, A.; PARABONI, I. Searching brazilian twitter for signs of mental health issues. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2020. p. 6111–6117. Disponível em: <https://www.aclweb.org/anthology/2020.lrec-1.750>. Citado 16 vezes nas páginas 3, 9, 12, 18, 21, 24, 25, 26, 30, 31, 33, 34, 35, 36, 37 e 43.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear). [S.l.: s.n.], 2020. Citado na página 29.

YADAV, S. et al. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. p. 696–709. Disponível em: <https://aclanthology.org/2020.coling-main.61>. Citado 2 vezes nas páginas 20 e 21.

YAZDAVAR, A. H. et al. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. [S.l.: s.n.], 2017. p. 1191–1198. Citado 7 vezes nas páginas 9, 12, 20, 21, 26, 38 e 43.