Mental Health Prediction from Social Media Text Using Mixture of Experts

Wesley Ramos dos Santos Sungwon Yoon and Ivandré Paraboni

***Abstract*—Predicting mental health statuses from social media text is a well-known Natural Language Processing (NLP) task. In this work, we focus on the issue of depression and anxiety disorder prediction from Twitter by comparing a more conventional approach based on engineered features with a data-oriented alternative based on mixture of specialists with transformer language models. Results from a large corpus of depression/anxiety self-disclosed diagnoses in the Portuguese language are reported, and a feature importance analysis is carried out to provide further insights into these tasks.**

***Index Terms*—Natural Language Processing, Text Classification, Mental Health, Depression, Anxiety Disorder.**

I. Introdução

Transtornos de saúde mental como depressão e ansiedade são desafios bem conhecidos e uma crescente fonte de preocupação na sociedade moderna. Nesse contexto, as pesquisas em Processamento de Línguas Naturais (PLN) e áreas relacionadas têm procurado explorar a forma como esses transtornos podem estar refletidos na língua humana (por exemplo, tal qual é empregada nas redes sociais) propondo modelos computacionais capazes de detectá-los, preferencialmente antes do seu agravamento. Exemplos de iniciativas nesse sentido incluem a detecção automática de transtornos como depressão, ansiedade, automutilação, bipolaridade, vício em jogos, ideação suicida, anorexia e outros [1]–[7]. Estudos desse tipo, que procuram identificar casos de maior gravidade e eventualmente sinalizar a necessidade de um indivíduo buscar ajuda, são também o foco desse trabalho.

A tarefa computacional de predição de transtornos de saúde mental com base na linguagem empregada em redes sociais é um problema de pesquisa de interesse na área de PLN, sendo na maioria dos casos tratado como uma tarefa de aprendizado de máquina (AM) do tipo supervisionado, ou seja, fazendo uso de córpus de publicações (e.g., *tweets*) rotuladas com informações relativas ao estado mental dos usuários que as produziram para fins de treinamento e teste de modelos de classificação desses transtornos.

O presente trabalho aborda a questão da detecção de indivíduos com maior risco de virem a ser futuramente diagnosticados com transtornos de depressão e ansiedade a partir de dados textuais disponibilizados na rede social Twitter em Português. De forma mais específica, consideramos para esta tarefa o uso de modelos baseados em *mistura de especialistas* (MoE) [8], os quais têm obtido considerável sucesso em tarefas relacionadas à saúde mental em estudos recentes [9], [10]. Modelos desse tipo, que são baseados na combinação de especialistas mediados por uma estrutura auxiliar denominada rede *gating*, foram comparados ao uso de modelos convencionais baseados em engenharia de características mais propícias à interpretação humana. As principais contribuições previstas nesse estudo são as seguintes:

- Apresentação de modelos de mistura de especialistas de alto desempenho na tarefa de predição de transtornos de depressão e ansiedade a partir de textos em Português.
- Apresentação de modelos baseados em engenharia de características para as mesmas tarefas.
- Análise comparativa e interpretação dos modelo baseados em engenharia de características.

School of Arts, Sciences and Humanities, University of São Paulo e-mail:{wesley.ramos.santos,sungwon.yoon,ivandre}@usp.br.

II. Conceitos básicos

A tarefa computacional de predição de transtornos de depressão/ansiedade é tipicamente implementada com uso de métodos supervisionados de aprendizado de máquina [11]. Em modelos desse tipo, a classe dita positiva é formada pelo conjunto de dados (e.g., *tweets*) produzidos por indivíduos reconhecidamente diagnosticados com o transtorno em questão. Esses indivíduos podem ser identificados, por exemplo, por meio de validação externa, como avaliação psiquiátrica, o uso de inventários de saúde mental e outros métodos. Apesar de confiável, entretanto, o emprego de validação externa tende a ter custo elevado, o que na prática pode limitar o volume de instâncias (ou casos de depressão/ansiedade) disponíveis para a construção e validação do modelo.

Em virtude dessas dificuldades, estudos da área de PLN frequentemente substituem a validação externa por métodos de seleção baseados em *autorrelatos* [11] como em '*Minha psicóloga acaba de anunciar que tenho transtorno de ansiedade*'. Autorrelatos desse tipo, que são comuns em redes sociais, não fazem parte do conjunto de dados a ser coletado (pois isso tornaria a tarefa de predição trivial), servindo entretanto como instrumento de identificação dos usuários de interesse com custo substancialmente inferior ao de uma validação externa. Assim, embora mais sensível a imprecisões ou falsidade (que de qualquer forma seriam possíveis também no caso de validação externa como um acompanhamento psiquiátrico etc.), é geralmente aceito que o maior volume de dados obtido tende a compensar o ruído adicional [11].

O uso de autorrelatos para identificação de instâncias positivas do problema (e.g., indivíduos depressivos) é a abordagem possivelmente mais comum na área de PLN [1], [5], [12], [13]. Entretanto, é importante observar que autorrelatos apresentam evidência apenas da *presença* de um transtorno, mas não da

sua *ausência*, e por isso não podem ser empregados para seleção dos indivíduos da classe negativa (e.g., indivíduos não depressivos). Em vez disso, modelos preditivos baseados em autorrelatos utilizam uma formulação particular do problema, na qual a classe negativa é modelada como um grupo de controle contendo indivíduos selecionados aleatoriamente. Esse grupo de controle (ou classe negativa), por ocultar um certo número de indivíduos portadores do transtorno de interesse (i.e., usuários selecionados aleatoriamente e que não mencionaram nenhum diagnóstico do tipo), é normalmente construído em proporção várias vezes superior ao tamanho da classe positiva [1], [12].

A predição de depressão/ansiedade com base em dados obtidos por meio de autorrelatos configura um problema de aprendizado fortemente desbalanceado, no qual o objetivo *não é* distinguir indivíduos com ou sem depressão[1]. Em vez disso, o objetivo passa a ser o de identificar aqueles indivíduos que possuem *um risco acima da média* de virem a receber um diagnóstico desse tipo. Esse risco, que entre os indivíduos da classe positiva será necessariamente muito próximo de 100%, deve ser muito mais baixo dentro do grupo de indivíduos selecionados aleatoriamente.

III. Trabalhos relacionados

Foi conduzido um levantamento de estudos de PLN voltados à predição de transtornos de depressão/ansiedade com base em dados textuais rotulados por meio de autorrelatos, sumarizados na Tabela I. Esses estudos são categorizados de acordo com o tipo de problema considerado (d=depressão, a=ansiedade, *=outros), conjunto de dados (R=Reddit, T=Twitter, B=Blogs), idioma (En=inglês, Pt=português), características da representação textual (e=embeddings, t=tópicos, n=informações de rede, u=informações referentes ao usuário, l=atributos psicolinguísticos LIWC, p=part-of-speech, d=dicionário de domínio, s=sentimentos/emoções, i=imagens, b=bag of words, m=metadados, h=informação temporal), e algoritmos ou métodos de aprendizado (e.g., CNN=redes neurais convolucionais, LSTM=long short-term neural networks, MLP=multilayer perceptron, LR= regressão logística, RF=Random Forest, DT=árvore de decisão etc.).

Com base nesse levantamento, observa-se que estudos de predição de depressão (d) tendem a ser mais frequentes do que aqueles focados em outros tipos de transtorno, como o de ansiedade (a). Em alguns casos, inclusive, ambos são tratados de forma combinada [16] ou considerando-se especificamente a questão da sua comorbidade [21].

Estudos existentes, com raras exceções, tendem a ser baseados em dados provenientes das redes sociais Twitter ou Reddit. Alguns projetos contemplam a construção de um conjunto de dados próprio, mas vários fazem uso de algum tipo de recurso já existente. Recursos desse tipo incluem córpus de publicações Reddit como RSDD [1], SMHD [5], eRisk [12], ou o córpus de publicações Twitter em [3]. Como esperado, a maioria dos estudos é baseada em conjuntos de dados em inglês (En).

[1]Ressaltando-se mais uma vez que a informação sobre a ausência de um determinado transtorno não existe nos dados disponíveis.

TABELA I
MODELOS COMPUTACIONAIS DE PREDIÇÃO DE DEPRESSÃO E ANSIEDADE

Ref.	Transtorno	Dados	Idioma	Modelo textual	Algoritmos
[1]	d	R	En	e	CNN
[3]	d	T	En	n, u, i, t, d, l, e	LR
[4]	a	R	En	e, t, l	MLP
[14]	a,*	T	En	s	VADER
[5]	d,a,*	R	En	b, e	FastText
[6]	d	R	En	e, p, m, t, d	CNN
[15]	d	B	Pt	s	DT
[16]	d,a	T	En	d, s, h	ensemble
[17]	d	R	En	s	SVM
[18]	d	R	En	h, m, n	RF
[19]	d	R	En	b	SS3
[7]	d	T	En	e, i	CNN
[20]	d	T	En	b, s, t, i, n, l, u	RF
[21]	d,a	R	En	e	LSTM
[9]	d,a	R	En	e	LSTM+CNN
[10]	d	R,T	En	e,s	LR+LSTM

Quanto aos tipos de representação textual utilizada, a Tabela I demonstra ainda que as abordagens existentes empregam uma ampla gama de informações provenientes da rede social (incluindo dados textuais, *timestamps*, relações de amizade, conexões e interações, informações demográficas e outras). Também quanto aos algoritmos e métodos computacionais, observa-se grande variedade de soluções, com maior presença de modelos neurais. No entanto, não foram identificados estudos que utilizam modelos de língua pré-treinados baseados em *transformers* nas tarefas em questão.

IV. Método proposto

Assim como em outras aplicações de AM, modelos de predição de transtornos de saúde mental a partir de dados textuais podem fazer uso de dois tipos de estratégia de modelagem de características de aprendizado: o uso de indicadores explícitos (e.g., sintomas de depressão) computados por meio de engenharia de características, ou o reconhecimento de padrões a partir de dados textuais em estado puro, ou seja, com pouco ou nenhum pré-processamento. A segunda estratégia, mais frequentemente associada a métodos de aprendizado profundo, tende a apresentar melhor resultados, mas modelos baseados em engenharias de características oferecem facilidades de interpretação humana que podem auxiliar seu desenvolvimento e aperfeiçoamento.

Com base nestas observações, o objetivo do presente estudo foi o de apresentar duas estratégias computacionais de predição de transtornos de depressão e ansiedade a partir de dados textuais (ou *timelines*) do Twitter em português, aqui modeladas como uma tarefa de classificação binária. A primeira estratégia, mais orientada à tarefa, é baseada em engenharia de características de motivação psicolinguística potencialmente relevantes para a solução do problema. A outra, puramente orientada a dados, é baseada em uma arquitetura neural do tipo mistura de especialistas (MoE) construídos com modelos de língua do tipo *transformer*. Estas alternativas são discutidas individualmente nas seções a seguir.

A. Modelo Baseado em Engenharia de Características

Uma forma convencional de modelar problemas como a predição de transtornos de depressão e ansiedade a partir de dados textuais é a engenharia de característica, ou seja, o uso de características de aprendizado projetadas especificamente para representar indicadores sabidamente correlacionados com transtornos de saúde mental. Características desse tipo, como os indicadores descritos no estudo em [22], incluem desde aspectos específicos do domínio (como a menção explícita a transtornos de saúde mental), aspectos linguísticos (e.g., o uso de pronomes de primeira pessoa, associado a um maior risco de depressão), psicológicos (e.g., emoções) e muitos outros, levando à construção de modelos ditos orientados à tarefa.

Embora modelos desse tipo costumem apresentar desempenho inferior ao de abordagens orientadas a dados (como o modelo MoE a ser discutido na próxima seção), o presente uso de engenharia de características possui duplo propósito: além de servir de sistema de *baseline* em nossos experimentos com MoE, um modelo de inspiração psicolinguística é muito mais facilmente interpretável, podendo assim propiciar maior entendimento da tarefa computacional em questão.

Nos experimentos realizados, foi utilizado um classificador simples do tipo regressão logística baseado no seguinte conjunto de características, cujos detalhes são discutidos no restante desta seção:

- *DA_Mentions*: proporção de tweets contendo os termos 'depressão' ou 'ansiedade'.
- *Meds_01*: proporção de tweets contendo termos de natureza médica em geral.
- *Symptoms* [1..9]: proporção de palavras associadas a sintomas de depressão.
- *LIWC* [1..64]: proporção de palavras associadas a diferentes categorias LIWC [23].
- *Abs*: proporção de termos absolutos no texto [24].
- *Night_01*: proporção de tweets publicados depois das 21h e antes das 6h da manhã, potencialmente capturando distúrbios de sono associados à depressão [25].
- *1pVerbs*: proporção de pronomes pessoais de primeira pessoa, obtidos por *PoS tagging*.
- *MeMimComigo*: proporção de palavras que correspondem a pronomes me/mim/comigo.
- *Gender*: o gênero linguístico (0=masculino, 1=feminino) mais frequentemente adotado pelo indivíduo.

Menções a transtornos de depressão/ansiedade em *DA_Mentions* e termos de natureza médica em *Meds_01* são formas diretas de capturar informação relativa ao problema em questão. Essas características são extraídas do texto de entrada a partir de palavras-chave observadas em uma porção de dados de desenvolvimento. As menções a depressão/ansiedade usam as próprias palavras de interesse. Os termos de natureza médica são referentes a eventos de diagnóstico ou tratamento para qualquer tipo de condição, menções a problemas de saúde mental, a medicação ou a um profissional da área de saúde mental.

Termos associados a sintomas usuais de depressão, representados pelas características *Symptoms*[1..9], foram computados a partir do léxico proposto em [2], traduzido para o português em [26]. Foram considerados nove indicativos desse tipo: falta de interesse, tristeza/humor depressivo, transtorno de sono, falta de energia, transtorno alimentar, baixa autoestima, problemas de concentração, hiperatividade/baixa atividade e pensamentos suicidas. Embora altamente indicativos dos fenômenos de interesse, é importante destacar que estas características não necessariamente estão presentes nos dados disponíveis (ou seja, um indivíduo não necessariamente usa o Twitter para discutir problemas desse tipo), e, portanto, seu alcance é consideravelmente limitado.

As 64 características *LIWC* são categorias lexicais de natureza psicolinguística descritas em [23], divididas em quatro grandes grupos: processos linguísticos (e.g., uso de pronomes, tempos verbais etc.), processos psicológicos (e.g., percepção, certeza etc.), língua falada (e.g., concordância etc.) e questões pessoais (e.g., trabalho, dinheiro, religião etc.). A versão do dicionário LIWC utilizada nesse trabalho contempla 64 categorias traduzidas para o português em [27]. Uma descrição detalhada das categorias individuais é apresentada em [23].

O estudo em [24] sugere que indivíduos com depressão tendem a se expressar utilizando mais termos absolutos como 'sempre', 'todo' etc. Assim, consideramos a característica *Abs* para representar termos desse tipo, fazendo uma adaptação da lista de termos originais em inglês em [24] e com acréscimo de expressões específicas da língua portuguesa como advérbios de sentido absoluto (e.g., 'totalmente', 'certamente' etc.). Estas expressões foram compiladas considerando-se palavras com mínimo de 80 ocorrências na porção de treinamento do córpus, e avaliadas por dois juízes.

Estudos como [25] sugerem que indivíduos com depressão tendem a ser mais ativos nas redes sociais no período noturno. Assim, a característica *Night_01* objetiva capturar a informação temporal das publicações.

Também em [22], sugere-se que o uso de pronomes de primeira pessoa (e.g., 'eu') é mais presente em indivíduos com depressão. Esta noção, que já é capturada por categorias específicas do léxico LIWC (como 'ipron' e 'i'), foi assim estendida de modo a contemplar duas peculiaridades da língua portuguesa: o uso de verbos em primeira pessoa mesmo diante da ausência do pronome 'eu' em *1pverbs*, e o uso de outras formas pronominais de primeira pessoa em *MeMimComigo*, considerando-se que ambos os casos podem ser indicativos de discurso centrado na pessoa do próprio autor.

Finalmente, consideramos ainda o gênero linguístico associado ao indivíduo em *Gender* que se refere ao gênero do complemento masculino/feminino mais frequentemente utilizado em construções do tipo 'eu sou' 'eu me sinto' e afins, e que é utilizado como estimativa do gênero com o qual o indivíduo se identifica em substituição à informação de gênero real, que não é disponibilizada pela rede social.

B. Mistura de Especialistas

Deixando-se momentaneamente de lado os modelos orientados à tarefa, e de forte inspiração psicolinguística e com alto grau de interpretabilidade, nesta seção discutimos uma abordagem puramente orientada a dados para o problema de predição de transtornos de saúde mental em texto.

Modelos de classificação baseados em redes neurais profundas são atualmente o estado-da-arte em problemas baseados em dados textuais como análise de sentimentos e na própria tarefa de predição de transtornos de saúde mental [9], [10]. Além disso, o uso de modelos de língua pré-treinados baseados em *transformer* como BERT [28] também tem apresentado ganhos significativos em muitas dessas tarefas. Com base nestas observações, propõe-se combinar modelos BERT em uma arquitetura de comitê de máquina dinâmico de mistura de especialistas [8] aqui denominada *MoE* [29].

A presente arquitetura *MoE*, baseada em [8], é constituída de três modelos especialistas com uma rede *gating* ponderadora. Cada modelo especialista, assim como sua rede *gating*, é constituído de um modelo de rede recorrente do tipo LSTM que recebe como entrada uma representação em nível de usuário construída pelo modelo de representação distribuída BERT [28]. A Figura 1 apresenta a representação de cada especialista e da rede *gating*.

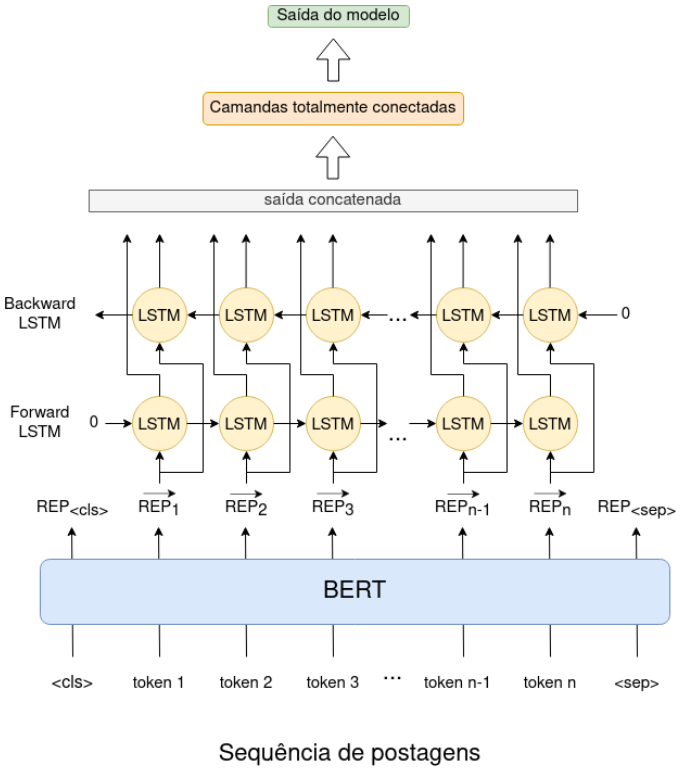


Fig. 1. Modelo BERT - adaptado de [30]

A Figura 2 apresenta a combinação de especialistas e rede *gating* da arquitetura proposta. O modelo *MoE* consiste de três redes especialistas (configuração esta que ofereceu o melhor equilíbrio entre acurácia e custo computacional) similares ao modelo apresentado na Figura 1. A rede *gating* também utiliza a mesma arquitetura, porém com uma camada de saída de tamanho três correspondente ao número de especialistas. A rede é atualizada com uso de Gradiente Descendente.

Todos os especialistas e a rede *gating* recebem as mesmas sequências consecutivas de 10 postagens cada como entrada. A classe de um indivíduo é dada por maioria de votos de todas as sequências que foram submetidas ao modelo. A rede *gating* fornece uma seleção por graduação, ou seja, cada especialista contribui de forma ponderada para a saída. Por questões de custo computacional, a construção desse modelo necessitou de compartilhamento de camadas de representação de modo que tanto modelos especialistas como a rede *gating* recebem a mesma representação BERT. No caso dos especialistas, a saída do modelo é binária, correspondendo às classes *Diagnosticados* ou *Controle*, enquanto na rede *gating* a saída é ternaria, correspondendo à ponderação dos três especialistas.

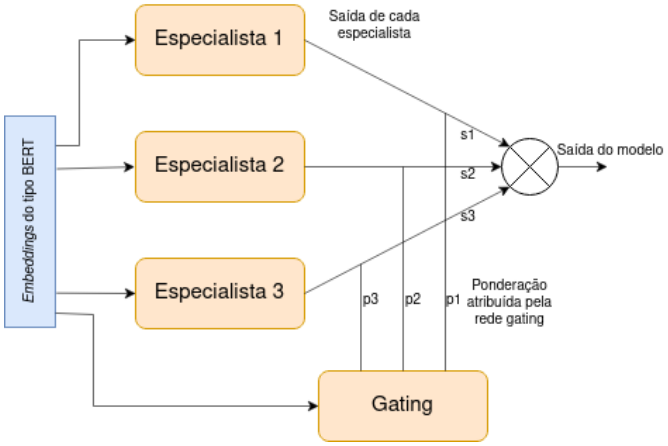


Fig. 2. Representação da arquitetura de mistura de especialistas

O uso de sequências consecutivas de *tweets* para representação de indivíduo a ser rotulado como *Diagnosticado* ou *Controle* é motivado pelo limite de 512 *tokens* de entrada do modelo BERT. A entrada da rede é uma sequência iniciada em uma posição aleatória da *timeline* a cada época. Os *tweets* são transformados em uma camada de representação distribuída do tipo BERT, que foi pré-treinado para o português em [31].

A saída da última camada desse modelo é utilizada como representação final da sequência de *tweets*. A arquitetura é seguida por uma rede BiLSTM de 4 camadas com 100 neurônios *backward* e 100 neurônios *forward* com função de ativação *ReLu*, que alimenta uma sequência de camadas totalmente conectadas (de tamanho 6000, 1000 e 100, respectivamente) com regularização do tipo *dropout* de 0,1, e utilizando *softmax* como função de ativação na camada de saída. Uma função binária de entropia cruzada com pesos de classe balanceados é definida para lidar com o desequilíbrio de classe, como segue:

$$E = \sum[-y * log(\widehat{y}) * W + (1-y) * (-log(1-\widehat{y}))]$$

Onde y é a saída esperada, $\widehat{y}$ é a saída do modelo, e W é o peso da classe positiva.

O modelo de mistura de especialistas é treinado com os 1.000 *tweets* mais recentes em cada *timeline*, divididos em 50 sequências de 10 tweets consecutivos iniciados aleatoriamente com possível sobreposição. Para o treinamento, a entrada corresponde às subsequências de 10 postagens consecutivas que podem ser selecionadas aleatoriamente em qualquer ponto da *timeline* a cada época (e portanto o tamanho total depende do número de épocas). Para a fase de predição, o conjunto de teste é formado a partir da seleção aleatória de 50 sequências de 10 postagens consecutivas iniciadas em um ponto aleatório

TABELA II
Timeline DE UM USUÁRIO QUE RELATOU UM DIAGNÓSTICO [MSG] EM UM MOMENTO ESPECÍFICO [END].

Data		Texto
25/03		Comprei um lanche mas esqueci em casa afff.
26/03		Nunca mais vou com ela ao cinema. Que mico.
28/03	[end]	LOL e eu achando que era na terça-feira
01/04		Quase chegando o aniversário e nada de notícia...
08/04		A vida é assim mesmo. Eu acho.
03/05		O jogo de ontem acabou com os meus nervos
20/05	[msg]	Mês passado a psicóloga me diagnosticou com depressão

da *timeline*, com possível sobreposição (e portanto o tamanho total pode ser inferior a 500 tweets). Para cada sequência, é atribuído um rótulo de *Diagnosticado* ou *Controle*. O rótulo do indivíduo é atribuído pela maioria dos rótulos de suas 50 sequências.

V. AVALIAÇÃO

Foram conduzidos experimentos de aprendizado de máquina supervisionado com o objetivo de comparar o desempenho dos modelos de engenharia de características e mistura de especialistas discutidos na seção anterior, aplicados às tarefas de predição de transtorno de depressão e ansiedade a partir de dados textuais da rede social Twitter em Português. Adicionalmente, os resultados do modelo de engenharia de características foram analisados detalhadamente como forma de fornecer uma interpretação acerca do seu funcionamento. Esses procedimentos e seus resultados são detalhados a seguir.

A. Conjunto de Dados

Os experimentos realizados fizeram uso de dados textuais disponibilizados pelo córpus SetembroBR [32], uma coleção de 47 milhões de *tweets* totalizando 555 milhões de palavras em Português, e que foram publicados por mais de 31 mil usuários. O córpus contempla dois tipos de dados textuais: conjuntos de *tweets* (ou *timelines*) públicos de indivíduos que relataram um diagnóstico de depressão/ansiedade, e um conjunto de controle formado por *timelines* de usuários aleatórios. Esses dois conjuntos, aqui denominados *Diagnosticados* e *Controle*, correspondem às classes positiva e negativa do problema de classificação, e são pareados por quantidade de publicações, gênero (masculino/feminino) e datas aproximadas de publicação. Além disso, seguindo-se práticas estabelecidas na área (e.g., [12]) discutidas na Seção II, a quantidade de usuários do conjunto *Controle* no córpus é sete vezes a quantidade de usuários do conjunto *Diagnosticados*.

No que diz respeito aos usuários *Diagnosticados*, o córpus contempla apenas a porção dita 'útil' (para fins de classificação) dos dados, ou seja, limitada aos *tweets* anteriores ao momento do diagnóstico feito (segundo o autorrelato publicado) por um especialista da área de saúde. Um exemplo de como esta porção de dados é delimitada é ilustrada pelo marcador [end] na Tabela II com base em um autorrelato indicado pelo marcador [msg].

Nesse exemplo, observa-se que, no dia 20 de maio (na parte inferior da *timeline*), o indivíduo menciona que teria recebido um diagnóstico no mês anterior (abril), em uma data não exata. Assim, as postagens que antecedem o mês de abril, até o item marcado como [end] no exemplo, são considerados como sendo o conjunto de dados de treino ou teste de modelos de predição de depressão. Todos os dados posteriores, a partir de 1o. de abril até o relato propriamente dito, marcado como [msg], são descartados de modo a não fazer parte do córpus.

O córpus contempla dois subconjuntos distintos de autorrelato — diagnóstico de depressão e de transtorno de ansiedade — a serem tratados como tarefas computacionais independentes. A Tabela III apresenta estatísticas descritivas dos dois conjuntos de dados utilizados.

TABELA III
ESTATÍSTICAS DESCRITIVAS DO CÓRPUS SETEMBROBR [32]

	Depressão		Ansiedade	
Estatísticas	Diagnosticados	Controle	Diagnosticados	Controle
Usuários	1684	11788	2219	15533
Tweets (milhões)	2,43	16,99	3,43	23,98
Palavras (milhões)	29,32	201,94	42,24	281,51

O córpus possui uma divisão aleatória padrão entre usuários de treino (80%) e teste (20%), a qual foi mantida nos experimentos conduzidos.

B. Resultados

Os modelos de engenharia de características e MoE foram treinados com base na porção de treino do córpus SetembroBR e testados com base na porção de teste do mesmo, medindo-se precisão (P), revocação (R) e medida F1 macro. A Tabela IV apresenta os resultados obtidos aplicados às tarefas de predição de transtorno de depressão e ansiedade.

TABELA IV
RESULTADOS GERAIS DE CLASSIFICAÇÃO. A MAIOR MEDIDA F1 DE CADA TAREFA É DESTACADA.

	Depressão			Ansiedade		
Modelo	P	R	F1	P	R	F1
Eng. caract.	0,58	0,67	0,56	0,56	0,64	0,53
MoE	0,64	0,67	**0,65**	0,59	0,62	**0,60**

Com base nesses resultados, observa-se que o modelo de mistura de especialistas apresentou desempenho superior em ambas as tarefas. A diferença em relação ao modelo baseado em engenharia de características é estatisticamente significativa tanto na tarefa de predição de depressão ($\chi = 98$, $p < 0,05$) como transtorno de ansiedade ($\chi = 36$, $p < 0,001$) de acordo com o teste de significância de McNemar. Além disso, cabe observar que esses resultados são superiores também aos reportados em [32], incluindo tanto modelos baseados em contagens de palavras (e.g., contagens TF-IDF) como de rotulação de sequências com uso de BERT e afins.

C. Análise

Embora de desempenho superior, modelos como o de mistura de especialistas discutido são baseados em características de difícil interpretação humana, como as *subwords* consideradas na representação BERT [28]. Por outro lado, o uso

de engenharia de características de motivação psicolinguística proporciona uma oportunidade mais direta de interpretação dos resultados. Assim, o presente modelo baseado em engenharia de características foi utilizado em dois tipos de análises complementares: na avaliação dos seus subconjuntos gerais de características, e para interpretação global do modelo baseada em importâncias de permutações. Estas duas análises são descritas individualmente a seguir.

A Tabela V apresenta os resultados de diferentes subconjuntos do modelo baseado em engenharia de características. Para uma definição dos conjuntos LIWC considerados, ver [23].

TABELA V
RESULTADOS DE CLASSIFICAÇÃO USANDO DIFERENTES SUBCONJUNTOS DO MODELO BASEADO EM ENGENHARIA DE CARACTERÍSTICAS.

	Depressão			Ansiedade		
Características	P	R	F1	P	R	F1
Todas	0,58	0,67	0,56	0,56	0,64	0,53
Depressão/ansiedade	0,57	0,61	0,57	0,58	0,63	0,58
Termos médicos	0,56	0,62	0,56	0,58	0,63	0,58
LIWC	0,56	0,63	0,51	0,54	0,60	0,49
Processos linguísticos	0,55	0,61	0,49	0,54	0,58	0,48
Processes psicológicos	0,54	0,59	0,48	0,53	0,58	0,47
Língua falada	0,52	0,54	0,43	0,51	0,52	0,42
Questões pessoais	0,51	0,54	0,43	0,51	0,53	0,44
Me, mim e comigo	0,54	0,59	0,49	0,52	0,54	0,46
Verbos de 1a. pessoa	0,54	0,59	0,47	0,54	0,58	0,47
Sintomas de depressão	0,53	0,57	0,51	0,53	0,57	0,51
Tweets noturnos	0,52	0,55	0,47	0,54	0,59	0,49
Média de tweets por dia	0,52	0,53	0,51	0,48	0,47	0,31
Termos absolutos	0,50	0,50	0,50	0,51	0,53	0,44
Gênero M/F	0,44	0,50	0,47	0,44	0,50	0,47

Com base nesses resultados, observa-se que há uma ligeira predominância das características que representam menções diretas a depressão/ansiedade e termos de natureza médica. De modo geral, entretanto, nenhum dos subconjuntos de características analisados se destaca dos demais de forma significativa, e os resultados relativamente próximos em todos os cenários considerados sugerem algumas das limitações comuns de modelos desse tipo, tipicamente baseados em um número reduzido de características e de alto custo de expansão.

Adicionalmente, a Tabela VI apresenta as características consideradas mais importantes de cada tarefa, computadas como importâncias de permutações [33], na qual os pesos associados a cada característica representam a melhoria no resultado da métrica de avaliação quando a característica especifica é substituída.

Os resultados desta análise sugerem, mais uma vez, a correlação entre o uso de termos de natureza médica (Meds_01) e menções explícitas a depressão/ansiedade (DA_Mentions) e esses transtornos. Além disso, observa-se uma certa consistência com os indicadores de transtornos de saúde mental em [22], como o discurso em primeira pessoa (liwc4_i e 1p_verbs), características humanas (liwc26_human) e processos cognitivos (liwc33_cogmech). A relativa semelhança entre o conjunto de características mais importantes de ambas tarefas é explicável pela observação de que parte dos indivíduos do córpus SetembroBR são diagnosticados com ambos os transtornos, e que ansiedade pode também ser um sintoma associado à depressão [34], o que dificulta a distinção completa entre os dois tipos de diagnóstico com base apenas em publicações de redes sociais.

TABELA VI
CARACTERÍSTICAS MAIS RELEVANTES PARA O MODELO COMPLETO

Depressão		Ansiedade	
Peso	Característica	Peso	Característica
+8.648	Meds_01	+9.281	Meds_01
+8.090	1pVerbs	+9.207	1pVerbs
+7.290	liwc4_i	+7.624	liwc4_i
+4.429	liwc3_ppron	+5.081	DA_Mentions
+4.248	DA_Mentions	+3.481	liwc26_humans
+3.592	liwc26_humans	+2.945	liwc18_conj
+2.836	liwc2_pronoun	+2.930	liwc3_ppron
+2.288	liwc18_conj	+2.727	liwc34_insight
+2.179	liwc34_insight	+2.591	liwc33_cogmech
+2.054	liwc23_social	+2.251	liwc2_pronoun

VI. CONSIDERAÇÕES FINAIS

Esse trabalho discutiu o uso de modelos baseados em engenharia de características de motivação psicolinguística orientados à tarefa, e modelos orientados a dados do tipo mistura de especialista BERT, em ambos os casos aplicados à tarefa de predição de transtornos de depressão e ansiedade a partir de autorrelatos em redes sociais em português. Os resultados obtidos sugerem que a mistura de especialistas apresenta resultados superiores em ambas as tarefas, embora o uso de engenharia de características proporcione a construção de modelos mais facilmente interpretáveis.

Apesar de modesta, a presente arquitetura de MoE sugere uma abordagem promissora para o problema em questão. De forma mais específica, observamos que é possível realizar diversos tipos de modificações em modelos desse tipo visando direcionar seu aprendizado para características específicas do texto, o que poderia levar a melhoria dos resultados. Por exemplo, cada modelo fraco ou especialista da arquitetura poderia fazer uso de mecanismos atenção [35], ser treinados para compor um conjunto de modelos especialistas em tarefas menores baseado em características LIWC ou outras, ou ainda ser tratado como um problema de classificação multitarefa [36]. Outra modificação possível ainda seria o uso de misturas de especialista hierárquica e aprendizado por maximização da esperança [37]. Iniciativas desta natureza são deixadas como oportunidades de trabalhos futuros.

AGRADECIMENTOS

Esse trabalho conta com apoio FAPESP # 2021/08213-0. Os autores agradecem ao Centro de Inteligência Artificial (C4AI-USP) e ao apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP # 2019/07665-4) e da IBM Corporation. O primeiro autor recebe apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 (# 88887.475847/2020-00).

REFERÊNCIAS

[1] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in *Proceedings of EMNLP-2017*, (Copenhagen, Denmark), pp. 2968–2978, Assoc for Comp Ling, 2017.

[2] A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, "Semi-supervised approach to monitoring clinical depressive symptoms in social media," in *IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pp. 1191–1198, 2017.

[3] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *26th International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 3838–3844, 2017.

[4] J. H. Shen and F. Rudzicz, "Detecting anxiety on Reddit," in *4th Workshop on Computational Linguistics and Clinical Psychology*, (Vancouver, Canada), pp. 58–65, Assoc for Comp Ling, 2017.

[5] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and v Goharian, "SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions," in *COLING-2018*, (Santa Fe, USA), pp. 1485–1497, Assoc for Comp Ling, 2018.

[6] M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[7] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, and H. Leung, *SenseMood: Depression Detection on Social Media*, pp. 407–411. New York, USA: Association for Computing Machinery, 2020.

[8] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[9] V. Souza, J. Nobre, and K. Becker, "A deep learning ensemble to classify anxiety, depression, and their comorbidity from texts of social networks," *Journal of Information and Data Management*, vol. 12, no. 3, pp. 306–325, 2021.

[10] L. Ansari and S. Ji, "Ensemble hybrid learning methods for automated depression detection," *IEEE Transactions on computational Social Systems*, 2022.

[11] G. Coppersmith, M. Dredze, C. Harman, H. Kristy, and M. Mitchell, "CLPsych 2015 Shared Task: Depression and PTSD on Twitter," in *2nd Workshop on Computational Linguistics and Clinical Psychology*, (Denver, USA), pp. 31–39, Assoc for Comp Ling, 2015.

[12] D. E. Losada, F. Crestani, and J. Parapar, "Overview of eRisk 2019 Early Risk Prediction on the Internet," in *LNCS 11696*, 2019.

[13] W. R. dos Santos, A. M. M. Funabashi, and I. Paraboni, "Searching Brazilian Twitter for signs of mental health issues," in *12th International Conference on Language Resources and Evaluation (LREC-2020)*, (Marseille, France), pp. 6113–6119, ELRA, 2020.

[14] K. Loveys, P. Crutchley, E. Wyatt, and G. Coppersmith, "Small but mighty: Affective micropatterns for quantifying mental health from social media language," in *CLPsych-2017*, pp. 85–95, 2017.

[15] R. Nascimento, P. Parreira, G. dos Santos, and G. P. Guedes, "Identificando sinais de comportamento depressivo em redes sociais," in *BraSNAM-2018*, (Porto Alegre, Brazil), SBC, 2018.

[16] A. Kumar, A. Sharma, and A. Arora, "Anxious depression prediction in real-time social data," in *Intl. Conf. on Advances in Engineering Science Management & Technology*, (Dehradun, India), 2019.

[17] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. M. y Gómez, "Detecting depression in social media using fine-grained emotions," in *NAACL-2019 Proceedings*, (Minneapolis, USA), pp. 1481–1486, Assoc for Comp Ling, 2019.

[18] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro, "Early detection of depression: Social network analysis and random forest techniques," *J Med Internet Res*, vol. 21, no. 6, p. e12554, 2019.

[19] S. G. Burdisso, M. Errecalde, and M. M. y Gómez, "t-SS3: a text classifier with dynamic n-grams for early risk detection over text streams," *Pattern Recognition Letters*, vol. 138, pp. 130–137, 2020.

[20] A. H. Yazdavar, M. S. Mahdavinejad, G. Bajaj, W. Romine, A. Sheth, A. H. Monadjemi, K. Thirunarayan, J. M. Meddar, A. Myers, J. Pathak, and P. Hitzler, "Multimodal mental health analysis in social media," *PLOS ONE*, vol. 15, no. 4, pp. 1–27, 2020.

[21] V. Souza, J. Nobre, and K. Becker, "Characterization of anxiety, depression, and their comorbidity from texts of social networks," in *SBBD-2020*, (Porto Alegre, Brazil), pp. 121–132, SBC, 2020.

[22] R. Trifu, B. Nemes, C. Bodea-Hategan, and D. Cozman, "Linguistic indicators of language in major depressive disorder (MDD). An evidence based research," *Journal of Evidence-Based Psychotherapies*, vol. 17, pp. 105–128, 03 2017.

[23] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Inquiry and Word Count: LIWC*. Mahwah, NJ: Lawrence Erlbaum, 2001.

[24] M. Al-Mosaiwi and T. Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Science*, vol. 6(4), pp. 529–542, 2018.

[25] M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *International AAAI Conference on Web and Social Media (ICWSM)*, AAAI, 2013.

[26] A. Mendes, R. Passador, and H. Caseli, "Identificando sintomas de depressão em postagens do twitter em português do brasil," in *STIL-2021*, (Porto Alegre, Brazil), pp. 162–171, SBC, 2021.

[27] P. P. Balage Filho, S. M. Aluísio, and T. Pardo, "An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis," in *STIL-2013*, (Fortaleza, Brazil), pp. 215–219, 2013.

[28] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019 Proceedings*, (Minneapolis, USA), pp. 4171–4186, 2019.

[29] "Available in: https://dadosabertos.bcb.gov.br/dataset/11-taxa-de-juros—selic,"

[30] A. S. Bock and I. Fine, "Anatomical and functional plasticity in early blind individuals and the mixture of experts architecture," *Frontiers in human neuroscience*, vol. 8, p. 971, 2014.

[31] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: pretrained BERT models for Brazilian Portuguese," in *9th Brazilian Conference on Intelligent Systems (BRACIS) - LNCS 12319*, (Cham), Springer, 2020.

[32] W. R. dos Santos, R. L. de Oliveira, and I. Paraboni, "SetembroBR: a social media corpus for depression and anxiety disorder prediction," *Language Resources and Evaluation*, 2023.

[33] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders 5th edition*. Arlington,USA: American Psychiatric Association, 2013.

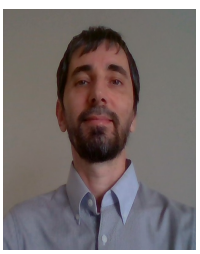
[35] Q. Cong, Z. Feng, F. Li, Y. Xiang, G. Rao, and C. Tao, "Xa-bilstm: A deep learning approach for depression detection in imbalanced data," in *BIBM-2018*, pp. 1624–1627, IEEE, 2018.

[36] A. Benton, M. Mitchell, and D. Hovy, "Multi-task learning for mental health using social media text," in *EACL-2017*, (Valencia, Spain), pp. 152–162, Assoc for Comp Ling, 2017.

[37] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994.

Wesley Ramos dos Santos Information Systems PhD candidate at the School of Arts, Sciences and Humanities (EACH), University of São Paulo (USP).

Sungwon Yoon Information Systems undergraduate student at the School of Arts, Sciences and Humanities (EACH), University of São Paulo (USP).



Ivandré Paraboni PhD in Computer Science (University of Brighton, UK), and associate professor at the School of Arts, Sciences and Humanities (EACH), University of São Paulo (USP).