Proposta recebida em Agosto 2024 e aceite para publicação em Novembro 2024.

Detecção precoce de transtornos de saúde mental em Português

Portuguese mental health early risk prediction

Bruno Issamo Tagava Nagamatu
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo

Ivandré Paraboni
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo

Resumo

O presente estudo enfoca o problema da detecção precoce de transtornos de saúde mental nos moldes da série de desafios *eRisk* (originalmente voltado ao domínio de fóruns de discussão sobre saúde mental no idioma inglês) em uma rede social de propósito geral em português. De forma mais específica, propõe a adaptação de uma estratégia vencedora em diversas edições desse *shared task* para o caso da detecção precoce de depressão e transtorno de ansiedade no domínio do Twitter/X brasileiro, usando para esse fim uma abordagem inédita baseada em LLMs com uso de engenharia de *prompts*. Os resultados obtidos indicam que o uso de LLMs apresenta maior poder de antecipação de diagnóstico em relação a abordagens tradicionais da área, e que a detecção com base em publicações da redes social de propósito geral é potencialmente mais desafiadora do que na formulação original do problema, sendo dependente da proximidade das mensagens do momento do diagnóstico na ordem cronológica da *timeline* do *Twitter*/X.

Palavras chave

saúde mental; depressão; ansiedade; detecção precoce; redes sociais

Abstract

The present study focuses on the early detection of mental health disorders along the lines of the eRisk shared task series (originally devoted to online mental health discussion domain in the English language) in general-purpose Portuguese social media. More specifically, we adapt a strategy that has won a number of shared tasks to the case of early detection of depression and anxiety disorders in the Brazilian Twitter/X domain, using for this purpose a novel approach based on LLMs and prompt engineering. Our results indicate that the use of LLMs affords greater power to anticipate diagnosis if compared to traditional approaches in the field, and that detection based on general-purpose social media text is potentially more challenging than in the original problem formulation, being dependent on the proximity of the messages to the moment of diagnosis in the chronological order of the timeline on Twitter/X.

Keywords

mental health; depression; anxiety; early risk prediction; social media

1. Introdução

Transtornos de saúde mental como depressão, ansiedade, bipolaridade, distúrbios alimentares, esquizofrenia e outros atingem cerca de 10,7% da população mundial (Dattani et al., 2021). Nesse contexto, o aumento da popularidade das mídias sociais fez com que esse meio se tornasse uma rica fonte de dados para pesquisadores da área de Processamento de Línguas Naturais (PLN). Em estudos desse tipo, textos provenientes de redes sociais são comumente utilizados para detecção de transtornos de saúde mental com uso de métodos de aprendizado de máquina supervisionados (Zafar & Chitnis, 2020; Chiong et al., 2021; Amanat et al., 2022; Chen et al., 2023; Lestandy et al., 2024), constituindo uma tarefa que pode ser vista como um caso particular de caracterização autoral ou *author profiling* (da Silva et al., 2020; Bevendorff et al., 2022; Flores et al., 2022; Pavan et al., 2023).

Para fins de treinamento e teste, modelos tradicionais de detecção de transtornos de saúde mental a partir de texto tipicamente fazem uso da totalidade dos dados (e.g., *timelines*) disponíveis na forma de um córpus rotulado. A tarefa computacional assume assim a forma de um problema de análise da rede social em que, dada uma *timeline* de interesse, o objetivo é decidir se aquele conjunto de publicações corresponde, por exemplo, a um perfil depressivo ou não.

Apesar das evidentes aplicações práticas de abordagens desse tipo, entretanto, estudos como

DOI: 10.21814/lm.16.2.440

os participantes da série de desafios (ou *shared tasks*) *eRisk* (Losada et al., 2017, 2018, 2019; Parapar et al., 2022, 2023) introduziram o conceito de detecção *precoce* de transtornos de saúde mental, na qual o objetivo passa a ser a detecção da forma mais antecipada possível, ou seja, com base no menor volume possível de publicações seguindo a ordem cronológica em que são apresentadas na rede social.

Aplicações voltadas à detecção precoce são mais associadas à prevenção de transtornos do que à detecção de transtornos já diagnosticados, o que pode apresentar certas vantagens de natureza prática. Por exemplo, podem agilizar o suporte a indivíduos que manifestam os primeiros sinais de um transtorno grave, ou até mesmo auxiliar na sua prevenção (Parapar et al., 2023). Entretanto, observa-se que as tarefas tratadas nos desafios *eRisk* são quase que exclusivamente baseadas em córpus do gênero *Reddit* (e.g., discussões *online* sobre saúde mental, ou outros tópicos específicos) cujos dados são de natureza distinta de uma rede social de propósito geral, ou seja, que discutem potencialmente qualquer tipo de assunto de forma irrestrita.

Um exemplo típico de rede social dita de propósito geral, e que é o objeto de estudo da presente pesquisa, é o *microblog* *Twitter*/X. Em redes sociais desse tipo, conforme será discutido na seção 4, observa-se que a maior parte das publicações tende a não ter relação direta com o estado de saúde mental do usuário autor destas publicações, o que configura um problema computacional bastante distinto do considerado na formulação original do problema da série de desafios *eRisk*, em que o conteúdo textual trata exclusivamente do transtorno a ser detectado (e.g, depressão).

Além disso, cabe observar também que os estudos da série *eRisk* são voltados ao idioma inglês, e apenas recentemente começam a ser exploradas em iniciativas correlatas desenvolvidas para a língua espanhola (Mármol-Romero et al., 2023). No caso específico da língua portuguesa, por outro lado, embora existam iniciativas baseadas na formulação tradicional do problema (Mann et al., 2020; de Souza et al., 2022; dos Santos et al., 2023; dos Santos & Paraboni, 2023, 2024) e outras (de Cássia Alves et al., 2024; Mendes & Caseli, 2024), não há, até onde temos conhecimento, estudos de detecção precoce de transtornos de saúde mental baseados em dados em português.

Finalmente, observa-se ainda que, pelo caráter competitivo de desafios como *eRisk* e similares, e suas restrição de tempo para entrega de resultados, estudos existentes tendem a privilegiar métodos computacionais mais tradicionais e de menor custo computacional em detrimento de alternativas mais robustas, porém com custo computacional geralmente superior. Esse é o caso, por exemplo, do uso de grandes modelos de língua (LLMs) que, embora tenham começado a ser aplicados à formulação tradicional do problema de detecção de transtornos de saúde mental (Tao et al., 2023; Agrawal, 2024), ainda não figuram entre as melhores soluções existentes para detecção precoce na série de desafios *eRisk* e estudos correlatos.

Com base nestas observações, o presente estudo enfoca o problema da detecção precoce de transtornos de saúde mental em uma rede social de propósito geral em português, utilizando para esse fim uma abordagem baseada em LLMs. De forma mais específica, propõe-se a adaptação da estratégia adotada em Burdisso et al. (2019), Loyola et al. (2021), Loyola et al. (2022) e Thompson et al. (2023), vencedora em diversas tarefas da série de desafios *eRisk* no domínio *R*eddit em inglês, para o caso da detecção precoce de depressão e transtorno de ansiedade em português no domínio do *Twitter*/X brasileiro, usando uma abordagem inédita baseada em LLMs com uso de engenharia de *prompts*.

Nesse cenário, o presente estudo propõe-se a investigar se o uso de uma abordagem baseada em LLM permite maior antecipação de diagnóstico em relação à abordagens tradicionais da área, e se a proximidade das mensagens do momento do diagnóstico na ordem cronológica da *timeline* do *Twitter*/X influencia o resultado do modelo.

As principais contribuições apresentadas nesse estudo são as seguintes:

- Um novo modelo de detecção precoce de depressão/ansiedade em português usando uma abordagem de *prompts* de LLMs, com resultados de antecipação de risco superiores aos de abordagens tradicionais.
- Adaptação do problema de detecção precoce para o caso de redes sociais de propósito geral, como o *Twitter*/X.
- Abordagem de detecção precoce para o caso específico de transtorno de ansiedade, tema que é possivelmente inédito na literatura da área.

O restante desse artigo está organizado da seguinte forma. A seção 2 fundamenta o problema computacional de detecção precoce de transtornos de saúde mental. A seção 3 apresenta um le-

vantamento dos trabalhos correlatos no contexto da série de desafios *eRisk* e iniciativas similares. A seção 4 descreve os modelos propostos nesse trabalho. A seção 5 descreve o córpus utilizado para fins de treino e teste desses modelos, o procedimento de avaliação e seus resultados. Finalmente, a seção 6 sumariza as contribuições apresentadas e destaca possíveis direções de pesquisa futura.

2. Detecção precoce

A detecção precoce (de transtornos de saúde mental ou outros) tal qual proposta nos desafios da série *eRisk* (Losada et al., 2017; Parapar et al., 2023) tem como componente central a métrica de avaliação *f1-latency* introduzida em Sadeque et al. (2018). De acordo com esta definição, para cada usuário $u \in U$ a ser avaliado, todas as suas mensagens são analisadas em ordem cronológica (i.e., conforme aparecem na rede social). Após avaliar k_u ($k_u \geq 1$) mensagens, é tomada então uma decisão binária $d_u \in \{0, 1\}$ atribuindo um rótulo $g_u \in \{0, 1\}$ ao usuário para indicar se é um caso de risco ou não.

Um componente-chave desse tipo de avaliação deve ser o atraso na detecção de verdadeiros positivos, pois se deseja que os sistemas detectem esses casos o quanto antes. Portanto, uma primeira e intuitiva medida de atraso (*latency*) pode ser definida da seguinte forma:

$$\text{latency}_{TP} = \text{median}\{k_u : u \in U, d_u = g_u = 1\}$$

A latência é assim calculada sobre os verdadeiros positivos (TP) detectados pelo sistema, avaliando-se o atraso na decisão com base na mediana de mensagens que o sistema teve que processar para detectar esses casos positivos. Com isso, obtém-se a métrica *f1-latency* proposta em Sadeque et al. (2018), que combina a eficácia da decisão (estimada com a medida *f1* padrão) e o atraso na decisão. O cálculo é realizado multiplicando-se *f1* por um fator de penalidade baseado no atraso médio. Mais especificamente, cada decisão individual (verdadeiro positivo) tomada após a leitura das mensagens k_u recebe a seguinte penalidade:

$$\text{penalty}(k_u) = -1 + \frac{2}{1 + exp^{-p(k_u - 1)}}$$

onde p é um parâmetro que determina a rapidez com que a penalidade deve aumentar. Em Sadeque et al. (2018), p foi definido de forma que a penalidade seja igual a 0,5 do número médio de postagens de um usuário. Nota-se que, para uma decisão logo após a primeira escrita, não há penalidade, ou seja, $penalty(1) = 0$. O fator de velocidade geral do sistema é calculado como:

$$\text{speed} = 1 - \text{median}\{\text{penalty}(k_u) : u \in U, d_u = g_u = 1\}$$

onde a velocidade é igual a 1 para um sistema cujos verdadeiros positivos são detectados logo na primeira escrita. O sistema lento, que (por exemplo) detecta verdadeiros positivos apenas após centenas de mensagens, receberá uma pontuação de velocidade próxima a 0.

Finalmente, a métrica *f1* ponderada pela latência é simplesmente

$$f_{1_{\text{latency}}} = f_1 \cdot \text{speed}$$

Observa-se assim que há uma relação inversa entre a medida *f1* padrão e antecipação, ou seja, quanto mais mensagens forem examinadas, maior tende a ser *f1*, porém menor a possibilidade de antecipação medida por *f1-latency*.

3. Trabalhos relacionados

Variações da medida *f1-latency* discutidas na seção anterior são um dos principais critérios empregados na avaliação dos sistemas participantes da série de desafios (ou *shared tasks*) *eRisk*, que em anos recentes abordaram diversas tarefas computacionais relacionadas a transtornos de saúde mental como depressão, anorexia e outros (Losada et al., 2017, 2018, 2019; Parapar et al., 2022, 2023). Estas tarefas são organizadas com base em um córpus (tipicamente no domínio *Reddit* em inglês) com rotulação binária (risco ou não risco) em nível de usuário. De forma análoga, o mais recente desafio *MentalRiskES* (Mármol-Romero et al., 2023) apresentou pela primeira vez tarefas desse tipo para a língua espanhola. Não há, até onde temos conhecimento, similares voltados ao idioma português.

A Tabela 1 sumariza as tarefas de interesse para o presente trabalho abordadas nesses eventos, com indicativo do tamanho da tarefa em número de instâncias de treino e teste dos córpus utilizados.

3.1. Abordagens selecionadas

Dado que a maioria dos estudos de interesse para a presente pesquisa encontra-se no âmbito da série de desafios *eRisk* dedicado ao idioma inglês, optou-se por realizar um levantamento desses estudos no período 2019-2023, selecionando-se os estudos que obtiveram melhores resultados publicados no contexto destas competições em tarefas

Referência	Idioma	Tarefas	Treino (k)	Teste (k)
Losada et al. (2018)	En	depressão anorexia	531 -	545 -
Losada et al. (2019)	En	anorexia automutilação	254 -	571 170
Losada et al. (2020)	En	automutilação	170	103
Parapar et al. (2021)	En	apostas automutilação	- 170	1129 103
Parapar et al. (2022)	En	Apostas depressão	1128 -	1029 723
Parapar et al. (2023)	En	apostas	1128	1129
Mármol-Romero et al. (2023)	Es	anorexia depressão	6 6	4 5

Tabela 1: Detecção precoce de transtornos de saúde mental em inglês (En) e espanhol (Es).

relacionados à saúde mental. Além desses estudos, foram incluídos também alguns outros trabalhos que não participaram dos desafios *eRisk*, mas que fizeram uso desses conjuntos de dados, conceitos ou métricas de avaliação aplicados à tarefa análoga de detecção precoce de ideação suicida.

A Tabela 2 apresenta um resumo dos estudos selecionados, categorizados conforme o tipo de tarefa abordada (A=anorexia, D=depressão, G=apostas, S=automutilação, SU=ideação suicida), tipo de *feature* textual empregado quando pertinente (emb=*embeddings*, BoW=*bag-of-words*, BoSE=*bag-of-sub-emotions*, DAN=rede de associações profundas, LIWC=*Linguistic Inquiry and Word Count* (Pennebaker et al., 2001), bigrams), e principal método computacional de cada abordagem (SVM=*support vector machine*, SS3=*Smoothness, Significance and Sanction*, MLP=*multilayer perceptron*, ROBERTa, BERT (Devlin et al., 2019), simil.=similaridade textual, LSTM=*long short-term memory networks*, CNN=*convolutional neural networks*, LR=*logistic regression*, RNN=*recurrent neural network*, *Longformer* ou *Ensemble*). Finalmente, é apresentada ainda a posição no *ranking* de resultados da respectiva competição *eRisk*, quando aplicável. Nesta representação, modelos textuais e métodos marcados como '-' indicam que a proposta engloba representação textual e aprendizado em uma solução única (por exemplo, fazendo refinamento de um modelo BERT sem classificação adicional), ou que o trabalho selecionado não foi totalmente claro a esse respeito.

Dado que vários dos sistemas aqui relacionados foram reutilizados em anos subsequentes e em outras tarefas, com ou sem variações significativas em relação à participação inicial, a seguir discutimos apenas os melhores colocados de cada edição da série *eRisk* tomando por base na métrica de avaliação *f1-latency*, detalhada na seção anterior.

O estudo por Mohammadi et al. (2019), que obteve a melhor *f1-latency* da competição *eRisk* em 2019 para a tarefa de detecção de anorexia, propôs uma abordagem comum a diversos outros sistemas participantes da série de desafios. Nessa proposta, é utilizado um conjunto de submodelos neurais baseados em atenção para extrair características e prever probabilidades de classe, posteriormente usadas como entrada para um classificador do tipo SVM.

Na tarefa de detecção de automutilação no mesmo ano, o melhor resultado em termos de *f1-latency* foi obtido pelo estudo em Burdisso et al. (2019). Esse estudo definiu as diretrizes básicas de um modelo que seria reutilizado em tarefas similares nos anos subsequentes da competição, e que tinha como objetivo encontrar um balanço adequado entre a necessidade de maximizar a precisão na identificação de risco e, ao mesmo tempo, minimizar o tempo de uma detecção confiável. Para esse fim, foi proposta uma arquitetura baseada em duas etapas: classificação com informação parcial (*Classification with Partial Information* - CPI) e decisão do momento de classificação (*Deciding the Moment of the Classification* - DMC). Na primeira etapa desta abordagem, somente a classe representando risco é considerada, ou seja, se a entrada parcial for classificada como não-risco, o modelo segue

Domínio	Estudo	Tarefas	Texto	Método	Posição
eRisk2019	Mohammadi et al. (2019)	A	emb.	SVM	1
	Burdisso et al. (2019)	S,A	BoW	SS3	1,5
	Ragheb et al. (2019)	S,A	emb.	MLP	4,2
	Naderi et al. (2019)	S,A	BoW	SVM	2,7
	Aragón et al. (2019)	A	BoSE	SVM	3
	Trifan & Oliveira (2019)	S	BoW	SVM	3,9
	Masood et al. (2019)	A	BoW	SVM	4
eRisk2020	Martínez-Castano et al. (2020)	S	-	ROBERTa	1
	Bagherzadeh et al. (2020)	S	emb.	SVM	3
	Aragón et al. (2020)	S	BoW,BoSE	SVM	5
eRisk2021	Loyola et al. (2021)	G,S	BoW	SVM,SS3	1,1
	Bucur et al. (2021)	G,S	-	BERT	2,3
	Campillo-Ageitos et al. (2021)	S	BoW	SVM	3
	Maupomé et al. (2021)	G,S	emb.	simil.,RoBERTa	4,8
	Lopes (2021)	G,S	emb.	LSTM,CNN	5,11
eRisk2022	Fabregat et al. (2022)	G	emb.	simil.	1
	Srivastava et al. (2022)	G,D	-,BoW	Longformer,SVM	7,1
	Mármol-Romero et al. (2022)	G	-	RoBERTa	2
	Bucur et al. (2022)	G,D	-	RoBERTa	4,2
	Loyola et al. (2022)	G,D	BoW,-	SVM,SS3	3,3
	Sauberli et al. (2022)	D	-	BERT,LR	4
	Ferreira et al. (2022)	G,D	BoW	Ensemble	5,5
eRisk2023	Molina et al. (2023)	G	emb.	SVM	1
	Fabregat et al. (2023)	G	-	RNN	2
	Larrayoz et al. (2023)	G	DAN	FFNN	3
	Thompson et al. (2023)	G	-	BERT,SS3	4
outros	Gollapalli et al. (2021)	SU	emb.	LSTM	-
	Gamoran et al. (2021)	SU	LIWC	LR	-
	Morales et al. (2021)	SU	BoW	Ensemble	-
	Wang et al. (2021)	SU	emb.	SVM	-
	Bayram & Benhiba (2021)	SU	bigrams	Ensemble	-

Tabela 2: Trabalhos relacionados

apenas acumulando mais informações. Na segunda etapa, é aplicada uma política que decide quando parar de ler as entradas e classificar o usuário em questão como sendo de risco caso indicado na etapa anterior. Em versões posteriores desta abordagem, apresentadas em Loyola et al. (2021), Loyola et al. (2022) e Thompson et al. (2023), diversos tipos de representação textual (e.g., *bag-of-words*, *embeddings*, etc.) e métodos de classificação (e.g., SVMs, modelos neurais, etc.) foram experimentados, porém mantendo-se a arquitetura principal baseada nas etapas CPI e DMC. Na competição de 2021, em especial, a abordagem de Loyola et al. (2021) obteve o melhor resultado tanto na tarefa de detecção de vício patológico em apostas quanto na tarefa de detecção de automutilação.

O estudo de Martínez-Castano et al. (2020) obteve o melhor resultado em termos de *f1-latency* para a tarefa de detecção de automutilação do *eRisk* 2020. A abordagem proposta utilizou um modelo do tipo XLM-RoBERTa treinado em um conjunto de textos produzidos por indivíduos da classe positiva (i.e., automutilação), e textos produzidos por indivíduos aleatórios. Dado um conjunto de postagens de teste, a classe a ser atribuída é calculada com base na probabilidade de pertencer a cada um dos dois grupos, utilizando-se de parâmetros ajustados para refletir o limiar de probabilidade média mínima (θ), e as quantidades mínimas e máximas de textos necessários para classificação final (risco ou não risco).

O estudo por Fabregat et al. (2022) utilizou uma abordagem de vizinhos mais próximos para detecção precoce de vício em apostas, obtendo o melhor resultado de *f1-latency* para esta tarefa no contexto do desafio *eRisk* 2022. Nesta abordagem, uma mensagem é classificada como risco se os 20 vizinhos mais próximos também correspondem a mensagens consideradas de risco. Além disso, foi adotada uma estratégia de recriação dos rótulos do córpus de treinamento de modo a gerar uma rotulagem em nível de mensagem (e não em nível de usuário, que é o padrão nas tarefas *eRisk* aqui discutidas). Para esse fim, todas as mensagens de entrada inicialmente recebem o mesmo rótulo do seu autor (usuário), e são comparadas a subconjuntos de mensagens semelhantes, recebendo o rótulo do conjunto mais próximo. Nesse processo, assume-se que usuários da casse positiva podem conter mensagens negativas mas não o contrário, pois se usuários da classe negativa pudessem apresentar mensagens de teor positivo, então seriam rotulados como positivos.

No caso da tarefa de detecção precoce de depressão da mesma competição, o melhor resultado de *f1-latency* foi apresentado pelo estudo em Srivastava et al. (2022). Nesta abordagem, utilizaram-se diferentes métodos de engenharia de características e técnicas de classificação de texto. Na tarefa de predição de depressão, optou-se pelo uso de um modelo do tipo *bag-of-words* com ponderação tf-idf baseada na medida de entropia, em conjunto com um classificador SVM.

Finalmente, o estudo em Molina et al. (2023) também utilizou uma abordagem tradicional de aprendizado de máquina baseada em SVM com um modelo de contagens tf-idf de unigramas. Esta abordagem foi aplicada à detecção de vício patológico em apostas na competição *eRisk* 2023, obtendo o melhor resultado para a tarefa.

3.2. Considerações

Com base no levantamento realizado, observa-se que os estudos mais bem-sucedidos da área na série de desafios *eRisk* tendem a apresentar soluções computacionais relativamente simples se comparadas ao estado da arte de outras tarefas do PLN. Embora métodos mais atuais tenham sido recorrentes em todas as edições da série, as abordagens tradicionais apresentaram, de modo geral, resultados mais robustos. Uma possível explicação para esse cenário pode estar na própria natureza de eventos desse tipo, que tendem a privilegiar a entrega rápida de resultados com menor margem para a exploração de métodos de custo computacional elevado. Assim, métodos mais contemporâneos, como o uso de grandes modelos de língua (LLMs) como pretendido no presente trabalho, permanecem em grande parte abertos à investigação.

De especial interesse para o presente trabalho, são dignos de nota os resultados expressivos da arquitetura de duas etapas aplicada a diversas tarefas em Burdisso et al. (2019), Loyola et al. (2021), Loyola et al. (2022) e Thompson et al. (2023). Esta arquitetura será adaptada às tarefas de detecção de depressão e transtorno de ansiedade representadas pelo córpus *SetembroBR*, e utilizada tanto como ponto de partida para a presente estratégia – baseada em LLMs – como para a definição de modelos de *baseline*.

4. Detecção precoce de depressão/ansiedade em português

Esta seção aborda os problemas da detecção precoce de depressão e transtorno de ansiedade representados no córpus SetembroBR (dos Santos et al., 2023) com uso de LLMs, e a questão de como esta formulação do problema (baseado na rede social *Twitter*/X em português) se compara ao originalmente considerado no domínio *Reddit* na série de desafios *eRisk* (Losada et al., 2017) em inglês. Assim, o presente estudo propõe-se a investigar as seguintes questões de pesquisa:

Q1 O uso de uma abordagem baseada em LLM para detecção precoce de risco de transtorno de saúde mental permite maior antecipação de diagnóstico em relação à abordagens tradicionais da área?

Q2 A proximidade das mensagens do momento do diagnóstico na ordem cronológica da *timeline* do *Twitter*/X influencia a tarefa de detecção precoce de transtornos de saúde mental?

A questão de pesquisa $Q1$ é motivada pelo interesse em empregar métodos computacionais mais recentes à presente tarefa — baseadas no uso de LLMs — tendo em vista uma possível melhoria nos resultados. Entretanto, mesmo sendo esperado algum incremento na acurácia da classificação textual com o uso de um modelo mais sofisticado desse tipo, o principal interesse desta investigação, assim como na série de desafios *eRisk* (Parapar et al., 2023), está na questão da antecipação do diagnóstico, ou seja, na capacidade do modelo realizar a classificação com base no menor volume de dados (i.e., postagens) possível.

A questão $Q2$ é motivada pela observação de que, ao contrário dos grupos de discussão sobre saúde mental da plataforma *Reddit* empre-

gados na série de desafios *eRisk*, a rede social *Twitter*/X possui mensagens tratando dos mais diversos assuntos, e não apenas transtornos de saúde mental. Além disso, como estas mensagens são normalmente publicadas em longos períodos de tempo, é possível que boa parte dos dados disponíveis tenha pouca ou nenhuma relevância para a saúde mental. Considerando-se ainda que estudos como Eichstaedt et al. (2018) sugerem que a porção de dados relevantes para a tarefa estaria concentrada dentro dos 3 meses de atividade mais recente na rede social, o foco desta segunda investigação é o efeito da ordenação temporal de mensagens – ou proximidade do momento do diagnóstico – presente no *Twitter*/X, porém ausente no domínio da série de desafios *eRisk* (Losada et al., 2017, 2018, 2019, 2020; Parapar et al., 2021, 2022, 2023).

4.1. Modelos desenvolvidos

Para investigação das questões de pesquisa $Q1$ e $Q2$ introduzidas na seção anterior, foi desenvolvida uma abordagem baseada em predições obtidas com auxílio do modelo GPT 3.5[1], aqui denominado *GPTrelev*, e duas alternativas de *baseline* denominadas *W2V.logreg* e *Soft.BERT* que são adaptações para o domínio do *Twitter*/X em Português do modelo vencedor da série de desafios *eRisk* 2022 (Parapar et al., 2022) apresentado em Loyola et al. (2022). O código da implementação desses modelos está publicamente disponível[2].

Tanto o modelo *GPTrelev* proposto como os *baselines* *W2V.logreg* e *Soft.BERT* seguem a mesma arquitetura geral proposta de Loyola et al. (2022). Esta arquitetura conta com dois módulos principais: um módulo de classificação parcial (CPI), e outro de decisão do momento da classificação (DMC) responsáveis, respectivamente, pela classificação em nível de postagens e pela consolidação destas classificações em uma predição final em nível de usuário (ou *timeline*).

O módulo CPI realiza a leitura e classificação dos dados de entrada em lotes de m mensagens (*tweets*), atribuindo um rótulo binário (diagnosticado ou controle) a cada grupo. Com base nos grupos de mensagens assim classificadas, o módulo DMC procura então estimar, com a maior antecipação possível, se o indivíduo que as publicou possui ou não risco de vir a receber um diagnóstico de depressão/ansiedade. Assim, a decisão é tomada com base no menor volume possível de dados classificados pelo primeiro módulo.

[1] https://platform.openai.com/docs/models

[2] https://github.com/brunoissamo/mestrado_deteccao_precoce_redes_sociais

O módulo DMC utiliza uma adaptação da política de decisão proposta de Loyola et al. (2022). Conforme descrito na seção 3, o módulo DMC determina, a partir do resultado dado pelo módulo CPI, qual o momento adequado para interromper a leitura de *tweets* e de fato realizar a classificação do indivíduo como sendo do grupo de risco ou não. Se a probabilidade de diagnóstico for superior a um δ pré-definido, a leitura é interrompida e o indivíduo é classificado como tendo risco acima da população em geral de vir a obter um diagnóstico de depressão/ansiedade. Caso contrário, o sistema continua a leitura do próximo lote de m mensagens até o fim da *timeline*, podendo eventualmente ser classificado como pertencente ao grupo controle, ou seja, sem risco acima da média da população geral de vir a receber um diagnóstico desse tipo. Os detalhes específicos da adaptação desta política nos modelos propostos são discutidos nas seções a seguir.

4.1.1. Modelo GPTrelev

O modelo *GPTrelev* proposto combina a política de análise de risco proposta de Loyola et al. (2022) com predições obtidas com auxílio do modelo GPT 3.5. A escolha desse LLM foi motivada pelos resultados positivos observados em tarefas análogas de predição de estados de saúde mental com base no córpus *SetembroBR* como as discutidas em dos Santos & Paraboni (2024), dentre outras.

Em linhas gerais, o modelo *GPTrelev* substitui o módulo de análise das publicações de Loyola et al. (2022), originalmente composto de um classificador SVM e *word embeddings* computados com uso de *word2vec* (Mikolov et al., 2013), por predições obtidas diretamente do LLM. A arquitetura geral do modelo *GPTrelev* é ilustrada na Figura 1.

Conforme ilustrado na Figura 1, a principal diferença entre o modelo *GPTrelev* e a proposta em Loyola et al. (2022) é a substituição das probabilidades dadas pelo classificador SVM por escores indicativos de transtorno de saúde mental em uma escala de 0 a 10, em que 0 indica baixos indícios de risco, que no presente trabalho foram obtidas a partir do estudo em dos Santos & Paraboni (2024) . De forma mais específica, foi submetido ao modelo GPT um *prompt* solicitando que cada mensagem fosse analisada buscando indícios linguísticos que pudessem sugerir a presença de sintomas contínuos de depressão, atribuindo um escore de 0 a 10 que pode ser visto como grau de afinidade de cada mensagem com o

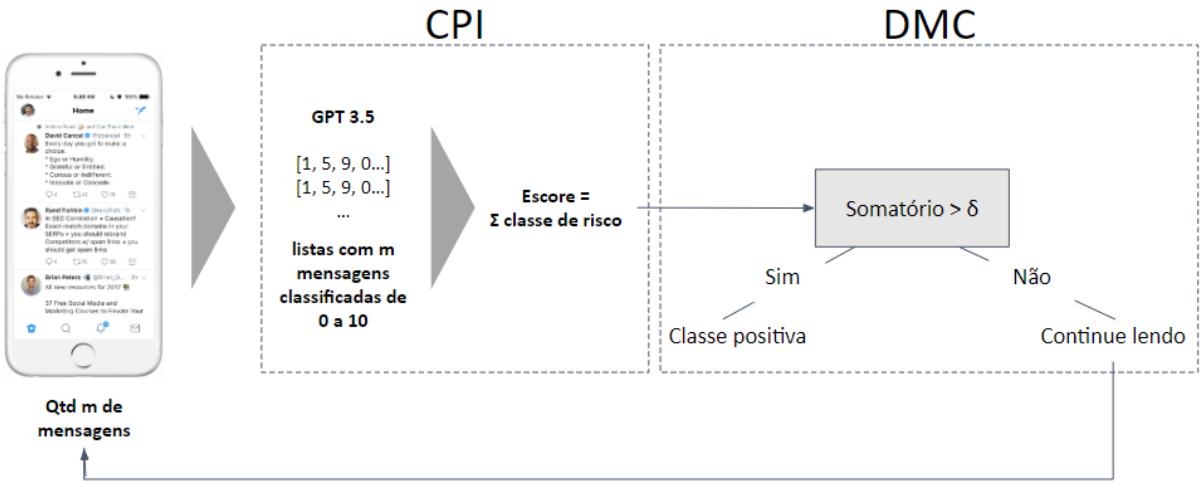

Figura 1: Arquitetura *GPTrelev*

tema de saúde mental, considerando expressões persistentes de desespero, tristeza profunda ou ansiedade constante como marcadores relevantes para realizar as classificações. A Tabela 3 sumariza a distribuição de escores obtidos sobre os conjuntos de teste do córpus *SetembroBR* para as tarefas de detecção de depressão e ansiedade.

	Depressão		Ansiedade	
Relevância	***Tweets***	%	***Tweets***	%
0	40.561	18,1	54.314	17,2
1	50.928	22,8	72.219	22,9
2	81.318	36,4	117.191	37,1
3	28.416	12,7	40.488	12,8
4	567	0,3	770	0,2
5	3	0,0	8	0,0
6	20.442	9,1	29.055	9,2
7	1.292	0,6	1.866	0,6
8	182	0,1	217	0,1
9	0	0,0	0	0,0
10	2	0,0	0	0,0

Tabela 3: Distribuição de escores de relevância para saúde mental produzidos pelo modelo GPT em dos Santos & Paraboni (2024)

Conforme observado na Tabela 3, o LLM de modo geral evita as categorias mais centrais, quase sempre comprometendo-se com uma resposta de baixa ou alta relevância. Além disso, observa-se que as categorias superiores (a partir de 8, e principalmente 9 e 10) praticamente nunca são selecionadas. O mesmo comportamento geral é observado no conjunto dos dados de ansiedade.

Na arquitetura *GPTrelev*, os escores de relevância gerados pelo LLM são somados em janelas de m mensagens, e esse resultado é fornecido ao módulo DMC (conforme Figura 1) para tomada de decisão. Caso esta soma seja maior que um valor δ, o usuário é classificado como diagnosticado. Do contrário, o sistema continua a leitura da próxima janela de mensagens. Foram testados diversos valores para os parâmetros δ e m, escolhendo aqueles que obtivessem maior valor da métrica *f1-latency* em um conjunto de validação do córpus. Para m foram testados os valores ímpares de 5 a 19 e para δ os valores $1, 5$, 2, $2,5$ e 3, sendo ao final utilizados $m = 15$ e $\delta = 2,5$. Cabe ressaltar que foi testado também o uso de janelas acumulativas dos dados e os resultados não foram expressivos, por isso optou-se por processar uma quantidade menor de informação, com janela móveis, e obter melhores métricas de desempenho.

4.1.2. Modelos de baseline

Como sistemas de *baseline* para comparação com o modelo GPTrelev descrito na seção anterior, foram consideradas duas abordagens baseadas na proposta original em Loyola et al. (2022), adaptadas ao domínio da rede social *Twitter*/X em português. Estas adaptações, que consistiram em encontrar novos valores para os parâmetros m e δ da abordagem original e usar *embeddings* para o português, foram necessárias em razão de diferenças na definição do problema de classificação (diagnosticados versus controle no córpus *SetembroBR* e depressivos versus não depressivos em Parapar et al. (2023)), gênero linguístico (rede social de propósito geral versus fórum de discussão sobre saúde mental), tamanho das publicações (*tweets* curtos versus postagens *Reddit* longas) e idioma (português versus inglês).

Assim como no modelo *GPTrelev* proposto, os dois modelos de *baseline* utilizados mantêm fixa a política de decisão, variando-se apenas a forma como as publicações individuais são avali-

adas. A arquitetura desses dois modelos é ilustrada na Figura 2 e detalhada a seguir.

Conforme ilustrado na Figura 2, para cada janela com m mensagens (*tweets*) o classificador do módulo CPI produz rótulos 0 (não diagnosticado) ou 1 (diagnosticado) utilizados na primeira etapa do DMC. Caso esse rótulo seja 0, o sistema continua lendo a próxima janela. Do contrário, verifica se a probabilidade obtida no primeiro módulo é maior que um valor pré-definido δ. Caso a probabilidade seja maior, o sistema interrompe a leitura e rotula o usuário como sendo diagnosticado. Caso contrário, continua lendo as mensagens até que finalize a *timeline*. O valor do parâmetro δ foi o mesmo utilizado no trabalho original em Loyola et al. (2022), como sendo 0,5.

Do ponto de vista da implementação, o modelo *W2V.logreg* faz uso de um classificador do tipo regressão logística e um modelo textual do tipo *bag-of-words*. Esta estratégia, embora não tenha sido a melhor combinação da competição *eRisk* 2022 (que usa um classificador do tipo SVM) apresentou os melhores resultados em estudos preliminares com o córpus *SetembroBR*.

Em complemento ao *baseline* *W2V.logreg*, considerou-se uma segunda alternativa, o modelo *Soft.BERT*, o qual também foi inspirado no modelo desenvolvido em Loyola et al. (2022), porém nesse caso substituindo-se a abordagem BoW com SVM por um classificador baseado no modelo BERT com saída *softmax*. Para adaptação ao contexto da língua portuguesa, realizou-se o *fine-tuning* na arquitetura bertabaporu-base-uncased (da Costa et al., 2023), um modelo BERT treinado em postagens do Twitter/X brasileiro como parte do projeto *SetembroBR*, e que inclui postagens produzidas pelos próprios usuários integrantes do córpus. Dado que em testes preliminares o uso de um número maior de épocas não revelou ganho significativo, por razões de custo computacional optou-se por realizar o *fine-tuning* com uso do otimizador AdamW por 3 épocas com taxa de aprendizado de 2e-5. Assim como no caso anterior, dada uma sequência de *tweets* de entrada obtém-se como resultado uma classificação e uma probabilidade, que será utilizado no módulo DMC, conforme ilustrado na Figura 2.

5. Avaliação

Como forma de investigar as questões de pesquisa $Q1$ e $Q2$ introduzidas na seção anterior, foram conduzidos dois experimentos com uso dos modelos *GPTrelev*, *W2V.logreg* e *Soft.BERT* descritos na seção 4.1. Nas seções a seguir é apresentado o conjunto de dados e procedimento de treino desses modelos, que é comum a ambos experimentos, e os resultados individuais de cada questão de pesquisa, seguidos de uma breve análise de erros.

5.1. Conjunto de dados

O presente trabalho é baseado no córpus *SetembroBR* apresentado em sua forma final em dos Santos et al. (2023) a partir do estudo-piloto em dos Santos et al. (2020). O córpus consiste de uma coleção de postagens (ou *tweets*) da rede social *Twitter*/X do Brasil de indivíduos diagnosticados com algum tipo de transtorno de depressão ou ansiedade, denominada classe *Diagnosticados*, e de uma classe *Controle* formado por indivíduos selecionados aleatoriamente, e que é de proporção sete vezes superior à classe *Diagnosticados*.

As *timelines* das classes *Diagnosticados* e *Controle* são pareadas por gênero (masculino e feminino), quantidade de publicações e período de publicação. Assim, para cada indivíduo da classe *Diagnosticado* há sete indivíduos da classe *Controle* com estas mesmas características. Nesta formulação do problema, amplamente adotada em diversos estudos da área (Coppersmith et al., 2015; Losada et al., 2017; Lynn et al., 2018; Cohan et al., 2018; Losada et al., 2019; Parapar et al., 2022), a tarefa computacional a ser modelada consiste em identificar indivíduos com probabilidade acima da média da população em geral (representada pela classe *Controle*) de vir a receber um futuro diagnóstico de depressão/ansiedade, e não se trata portanto de distinguir, por exemplo, indivíduos depressivos de não-depressivos.

No caso dos indivíduos diagnosticados, as publicações integrantes do córpus *SetembroBR* são limitadas ao período anterior ao momento em que o diagnóstico foi emitido por um profissional da área de saúde. O córpus é assim especialmente voltado ao desenvolvimento de aplicações de predição de depressão/ansiedade a partir do histórico de publicações que antecede o reconhecimento do transtorno e eventual tratamento.

A Tabela 4, adaptada de dos Santos et al. (2023), apresenta estatísticas descritivas da porção textual do córpus *SetembroBR* a ser utilizada no presente trabalho.

O córpus *SetembroBR* possui uma divisão aleatória padrão entre *timelines* de treinamento (80%) e teste (20%). A divisão existente será adotada também nos experimentos aqui relatados.

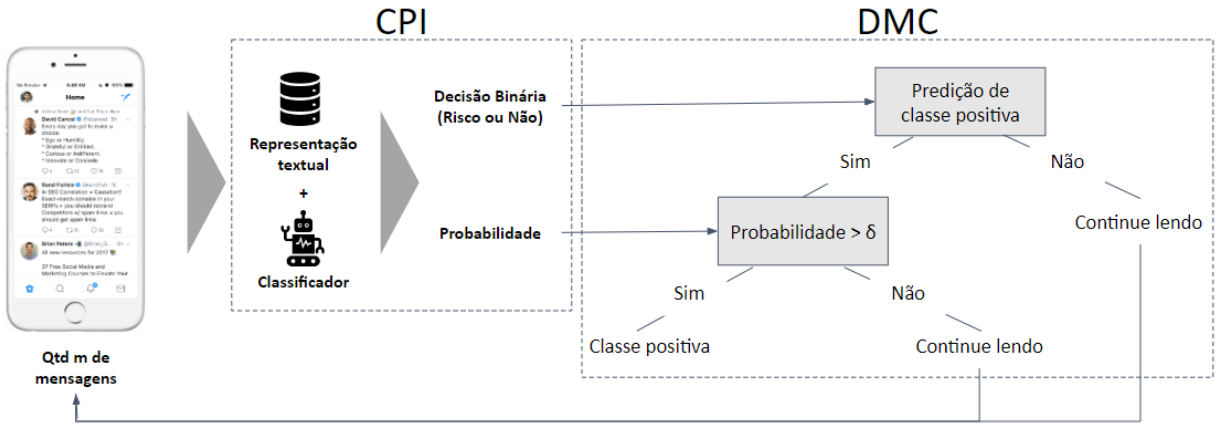

Figura 2: Descrição da Política de Decisão (PD) - *baselines*

	Depressão			Ansiedade		
Métrica	Diagnost.	Controle	Total	Diagnost.	Controle	Total
Usuários	1.684	11.788	13.472	2.219	15.533	17.752
Tweets (milhões)	2,43	16,99	19,42	3,43	23,98	27,41
Tokens (milhões)	29,32	201,94	231,26	42,24	281,51	323,75
Tweets/usuários	1.441	1.441	1.441	1.543	1.543	1.543
Tokens/tweets	12,08	11,88	11,91	12,33	11,74	11,81

Tabela 4: Estatísticas descritivas do córpus SetembroBR (dos Santos et al., 2023)

5.2. Procedimento

O treinamento e teste dos modelos *GPTrelev*, *W2V.logreg* e *Soft.BERT* fez uso das respectivas porções padrão de treino/teste do córpus *SetembroBR*. No caso do teste, entretanto, são usadas porções de dados diferentes para investigar as questões de pesquisa $Q1$ e $Q2$. Para a questão $Q1$, foram utilizadas apenas r mensagens mais recentes de cada *timelines* do conjunto de teste, sendo r o tamanho da menor *timeline* do córpus. Esta decisão é motivada pelo estudo em Eichstaedt et al. (2018), que sugere que os dados relevantes para fins de detecção de depressão concentram-se nos três meses mais recentes do histórico de mensagens de um indivíduo, e também por razões de custo computacional. Para a questão $Q2$, por outro lado, é utilizada a *timeline* completa de cada usuário.

O treinamento do modelo *GPTrelev* utilizou os rótulos de relevância para saúde mental produzidos pelo LLM para obter os parâmetros de decisão m e δ. Também por razões de custo, e dado que o córpus é de um gênero textual único e com certa sobreposição de dados (parte dos indivíduos diagnosticados com um transtorno é também diagnosticado com outro, e a classe Controle é parcialmente a mesma nos dois conjuntos), esses parâmetros foram posteriormente usados tanto para o classificador de depressão como ansiedade.

Seguindo Parapar et al. (2023), a avaliação foi realizada computando-se a medida *f1-latency*, que representa a ponderação entre o acerto e a quantidade de mensagens lidas. Além disso, dado que o córpus *SetembroBR* disponibiliza informação de data e hora das publicações, como forma de ilustrar o eventual ganho (i.e., antecipação de diagnóstico) obtido pelos modelos avaliados de forma possivelmente mais intuitiva, foi computada também a média de dias em que um diagnóstico poderia ser antecipado. Esse ganho é calculado como a média de dias transcorridos entre a data do último *tweet* da *timeline* e do último lido pelo sistema para cada usuário, dado em quantidade de dias. Finalmente computou-se também a medida *f1* da classe positiva como forma de destacar a relação inversa entre esta medida e *f1-latency*, acompanhada da medida *f1-macro* tradicional, a qual apresenta uma visão menos clara do desempenho do modelo na classe de interesse (ou seja, dos indivíduos diagnosticados) em virtude do maior tamanho da classe negativa (ou grupo Controle aleatório).

5.3. Experimento 1: o uso de LLM para detecção precoce de depressão/ansiedade

A Tabela 5 sumariza os resultados obtidos pelos modelos *GPTrelev*, *W2V.logreg* e *Soft.BERT* com

base na porção de teste do córpus *SetembroBR* no contexto da questão de pesquisa $Q1$, que trata do uso da abordagem baseada em LLM – o modelo *GPTrelev* – na detecção precoce de risco de depressão e transtorno de ansiedade.

Com relação à tarefa de detecção de depressão (na parte superior da Tabela 5), os resultados indicam que, de modo geral, o modelo *GPTrelev* é superior às abordagens de *baseline* de acordo com as métricas consideradas. Esta vantagem é evidente tanto no que diz respeito à antecipação do diagnóstico (representada pela medida *f1-latency*) como no que diz respeito à acurácia de classificação de diagnósticos (representada pela medida *f1-pos* da classe positiva). A exceção é observada no resultado de *f1-macro* do modelo *W2V.logreg*, que foi ainda ligeiramente superior ao resultado do modelo *GPTrelev*.

Com relação à tarefa de detecção de transtorno de ansiedade (parte inferior da Tabela 5), os resultados indicam novamente uma superioridade do modelo *GPTrelev*, embora o número de dias de antecipação do diagnóstico do modelo *Soft.BERT* tenha sido ligeiramente superior.

Considerando-se a questão da antecipação de risco representada pelas métricas *f1-latency* e número médio de dias de antecipação, observa-se tanto na detecção precoce de depressão como de transtorno de ansiedade uma divisão clara entre os modelos avaliados. Em um extremo, observa-se que os resultados de antecipação obtidos pelos modelos *Soft.BERT* e *GPTrelev* são relativamente próximos, enquanto que os do modelo *W2V.logreg* ocupam um lugar mais distante mesmo considerando-se que esta diferença não tenha tido grande reflexo em perda de acurácia (representada pelas medidas *f1-pos* e *f1-macro*). Assim, a adaptação do modelo *W2V.logreg*, que procura ser fiel à proposta original em Loyola et al. (2022), parece não ter sido tão bem sucedida no domínio *SetembroBR* quanto as alternativas aqui consideradas. Este resultado pode ter sido influenciado tanto pelas diferenças de gênero textual, de idioma e de definição da tarefa entre os córpus *eRisk* e *SetembroBR*, mas mais estudos são necessários para determinar porque estas diferenças são mais pronunciadas no uso de *embeddings* estáticos do tipo *word2vec*, e menor nas alternativas baseadas em BERT e GPT.

De forma conjunta, esses resultados oferecem uma resposta parcial à questão de pesquisa $Q1$, sugerindo que o uso de LLMs na forma proposta para detecção precoce de risco de transtorno de saúde mental permite maior antecipação de diagnóstico em relação a abordagens tradicionais desta área.

5.4. Experimento 2: o papel da informação temporal na predição precoce de depressão/ansiedade

Apesar da relativa superioridade do modelo baseado em LLM discutidos na seção anterior, os resultados obtidos para a tarefa de detecção precoce no córpus *SetembroBR* são claramente inferiores aos observados em tarefas análogas como as desenvolvidas pelos participantes dos desafios *eRisk* (Losada et al., 2017). Conforme discutido na seção anterior, esta diferença pode estar relacionada ao fato de que a natureza do problema – e dos dados – é distinta em pelo menos dois aspectos. Em primeiro lugar, há uma diferença de conteúdo a ser considerada. Ao contrário de grupos de discussão sobre saúde mental no *Reddit*, o córpus *SetembroBR* contém *timelines* tratando de qualquer tema. Em segundo lugar, há a noção de proximidade em relação ao momento do diagnóstico na ordem cronológica das *timelines* do *Twitter*/X, especialmente considerando-se que somente dados mais recentes são tidos como relevantes para esse fim em estudos como Eichstaedt et al. (2018).

Como forma de investigar estas diferenças, foi realizada uma análise comparativa dos resultados já apresentados na seção anterior, referentes ao modelo *GPTrelev* original (que recebe como entrada a porção da *timeline* de cada indivíduo de teste mais próxima possível do momento do diagnóstico), com resultados de uma versão modificada desse mesmo modelo, que recebe como entrada a *timeline* completa de cada indivíduo. O objetivo desta comparação foi o de verificar se, em uma situação prática em que o momento do diagnóstico não é conhecido (ou seja, tomando-se como entrada um conjunto de dados qualquer, que pode ou não conter mensagens de maior relevância para a detecção do transtorno) haveria perda de desempenho.

A Tabela 6 apresenta os resultados do modelo *GPTrelev* utilizando as postagens mais recentes (reproduzidos do experimento anterior para fins ilustrativos) e os resultados do mesmo modelo com base na *timeline* completa.

Os resultados apresentados na Tabela 6 mostram que, em linhas gerais, realizar a leitura sequencial de mensagens considerando-se apenas as mais próximas do momento do diagnóstico – como originalmente proposto no experimento da seção anterior – é realmente superior à leitura a partir da primeira mensagem disponível na *timeline*. Assim, ao contrário dos grupos de discussão *Reddit* da série *eRisk* (Parapar et al., 2023), a detecção precoce com base em *timelines* do *Twitter*/X é dependente do momento da

Tarefa	Modelo	f1-latency	Antecip. (dias)	f1-pos	f1-macro
Depressão	W2V.logreg	0,20	77	0,25	**0,59**
	Soft.BERT	0,20	**453**	0,22	0,14
	GPTrelev	**0,24**	436	**0,28**	0,56
Ansiedade	W2V.logreg	0,15	47	0,15	0,54
	Soft.BERT	0,21	**508**	0,22	0,47
	GPTrelev	**0,23**	479	**0,27**	**0,55**

Tabela 5: Q1: abordagem baseada em LLM GPTrelev e abordagens tradicionais. O melhor desempenho de acordo com cada métrica de avaliação é destacado.

Tarefa	Timeline	f1-latency	Antecip. (dias)	f1-pos	f1-macro
Depressão	Completa	0,13	296	0,27	0,39
	Mais próxima	**0,24**	**436**	**0,28**	**0,56**
Ansiedade	Completa	0,11	346	0,25	0,37
	Mais próxima	**0,23**	**479**	**0,27**	**0,55**

Tabela 6: Resultados do modelo GPTrelev com timeline completa e mais próxima do diagnóstico. O melhor desempenho de acordo com cada métrica de avaliação é destacado.

amostragem, o que responde de forma afirmativa à presente questão de pesquisa *Q*2. De forma mais específica, é provável que, em uma rede social de postagens ordenadas cronologicamente como no caso do *Twitter*/X, uma abordagem de detecção precoce apresentaria resultados ótimos apenas se os limites de informação temporal tipicamente considerados para fins de diagnóstico forem respeitados, como por exemplo o limite de três meses proposto em Eichstaedt et al. (2018). Este resultado é também consistente com as diretrizes de diagnóstico de depressão proposta na área de psicologia (American Psychiatric Association, 2013), que privilegiam o caráter recorrente e persistente de sintomas, e não sua ocorrência em episódios isolados.

5.5. Análise de erros

Como forma de identificar possíveis fontes de erro na abordagem *GPTrevel* proposta, foi realizada uma análise da distribuição das mensagens de diferentes graus de relevância para a tarefa utilizadas na presente abordagem, e que foram geradas com uso de um modelo GPT em dos Santos & Paraboni (2024). De forma mais específica, foram contabilizadas as quantidades de mensagens de grau de relevância baixa (escores de relevância 0, 1 e 2), média (escores 3, 4 e 5) e alta (6, 7 e 8) encontradas nas instâncias classificadas correta e incorretamente de cada classe (Diagnosticados ou Controle). Conforme discutido na seção 4.1, entretanto, é importante ressaltar que as mensagens de baixa relevância não fazem parte dos dados utilizados pelo modelo *GPTrevel*, sendo incluídas nesta análise apenas para fins comparativos.

A Tabela 7 apresenta os resultados obtidos para as tarefas de detecção de depressão e transtorno de ansiedade, com indicativo dos casos em que houve aumento (↑) ou diminuição (↓) do percentual de mensagens nas situações de classificação incorreta.

Tanto na tarefa de detecção de depressão como ansiedade, os resultados da Tabela 7 sugerem efeitos semelhantes. Nos casos incorretos da classe Diagnosticados com depressão/ansiedade, observa-se que as mensagens de média e alta relevância são *menos* frequentes. Inversamente, nos casos incorretos da classe Controle para depressão/ansiedade, as mensagens de média e alta relevância são *mais* frequentes.

Estes resultados indicam que a rotulação de relevância é uma possível fonte de erros de classificação tanto para a classe positiva como negativa, embora não seja possível afirmar a causa exata destas divergências. Uma possível explicação estaria no próprio projeto do córpus *SetembroBR*, cuja classe Controle naturalmente oculta um certo número de usuários aleatórios que podem ter depressão/ansiedade não relatada (dos Santos et al., 2023). Outra possibilidade, entretanto, é que a própria rotulação atribuída pelo modelo GPT em dos Santos & Paraboni (2024) apresente imprecisões.

	Depressão				Ansiedade			
	Diagnosticados		Controle		Diagnosticados		Controle	
Relev.	Corretos	Incorretos	Corretos	Incorretos	Corretos	Incorretos	Corretos	Incorretos
baixa	94,4%	95,33% ↑	98,68%	97,9% ↓	62,92%	65,76% ↑	81,45%	77,78% ↓
média	2,88%	2,67% ↓	0,98%	1,49% ↑	17,53%	17,32% ↓	11,42%	13,31% ↑
alta	2,71%	1,99% ↓	0,35%	0,62% ↑	19,56%	16,92% ↓	7,13%	8,90% ↑

Tabela 7: Distribuição de postagens de diferentes graus de relevância nas instâncias classificadas correta e incorretamente na detecção de depressão e ansiedade.

6. Considerações finais

Este artigo apresentou uma abordagem para detecção precoce de transtornos de saúde mental a partir de publicações em redes sociais baseada em LLMs, a qual foi comparada a abordagens tradicionais da área. Além disso, levando-se em conta a natureza das publicações em uma rede social de propósito geral, e sua distribuição ao longo de uma *timeline* de publicações, foi investigada também a adaptação do problema tal qual formulado na série de desafios *eRisk* (Parapar et al., 2023) – originalmente baseado em fóruns de discussão *Reddit* – para o domínio do *Twitter*/X brasileiro.

No que diz respeito à primeira questão – sobre o uso de LLMs na tarefa – os resultados obtidos sugerem que LLMs permitem maior antecipação do diagnóstico em relação à abordagem tradicionais, e que o uso de mensagens mais recentes demonstrou-se mais útil na detecção precoce de depressão e ansiedade.

Quanto à adaptação do problema de detecção precoce à rede social *Twitter*/X, observa-se que que há necessidade de filtrar mensagens de baixa relevância para a tarefa (e que são normalmente a maioria das mensagens em uma rede social de propósito geral), e levar em conta as publicações realizadas em momento mais próximo ao agravamento do problema (ou seus sintomas, etc.) já que, por exemplo, mesmo mensagens potencialmente relevantes (e.g., problemas como insônia) não são necessariamente indicativos de depressão se ocorrem de forma isolada ou distante do momento analisado (American Psychiatric Association, 2013).

Este estudo deixa uma série de oportunidades de melhoria. Em primeiro lugar, observa-se que os resultados, principalmente no que diz respeito à acurácia de classificação, são inferiores ao observados no domínio *eRisk*. Isso sugere que a adaptação ao domínio *Twitter*/X é uma tarefa complexa, e ainda não totalmente resolvida no presente trabalho. Uma possível explicação para esta diferença, e que carece de mais estudos, pode ser o fato de que, em uma rede social como o *Twitter*/X, as mensagens ditas relevantes para a tarefa são concentradas em pontos específicos da *timeline* (i.e., próximas ao momento em que o transtorno se agrava ou manifesta sintomas), o que possui impacto considerável sobre a capacidade do modelo de antecipar risco em uma leitura sequencial de mensagens ordenadas cronologicamente. Em outras palavras, dificuldades desse tipo podem não ser observadas, ou podem ser observadas em menor escala, na tarefa de detecção precoce baseada em fóruns de discussão especializada em saúde mental.

Finalmente, a própria questão do quê constitui uma mensagem relevante para a tarefa de detecção precoce de transtornos de saúde mental permanece em aberto. No presente trabalho, foi utilizado um método baseado em *prompts* submetidos ao modelo GPT para esta finalidade. Não está claro, entretanto, se esse método é realmente ideal para esta tarefa, ou como ele se compara a alternativas similares que seriam possíveis, por exemplo, utilizando outros LLMs. A investigação destas questões é também deixada como sugestão de trabalhos futuros.

Agradecimentos

Esse trabalho contou com apoio FAPESP #2021/08213-0.

Referências

Agrawal, Aryan. 2024. Illuminate: A novel approach for depression detection with explainable analysis and proactive therapy using prompt engineering. ArXiv [cs.CL]. 10.48550/arXiv.2402.05127

Amanat, Amna, Muhammad Rizwan, Abdul Rehman Javed, Maha Abdelhaq, Raed Alsaqour, Sharnil Pandya & Mueen Uddin. 2022. Deep learning for depression detection from textual data. *Electronics* 11(5). 676. 10.3390/electronics11050676