ANOTAÇÃO DE DADOS TEXTUAIS: UMA AVALIAÇÃO COMPARATIVA DE TRABALHOS BASEADOS NO USO DE LLM

LAERTY S. DA SILVA (IFPB, Campus João Pessoa), DAMIRES YLUSKA DE SOUZA (IFPB, João Pessoa)

E-mails: laerty.santos@academico.ifpb.edu.br, damires@ifpb.edu.br

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

Palavras-chave: llm; anotação de dados; depressão.

1. Introdução

A depressão é um transtorno mental global caracterizado por sintomas que variam desde baixa autoestima, desânimo e sensação de culpa até pensamentos suicidas (*American Psychiatric Association*, 2023). Textos em forma de autorrelatos postados em redes sociais ou fóruns online podem ser um meio para identificação de sinais da doença de modo precoce (Li et al., 2023). Para isso, modelos de aprendizado de máquina treinados a partir de exemplos podem ser utilizados. Entretanto, conjuntos de dados anotados para as quatro classes de depressão, conforme o Inventário de Beck (*Beck Depression Inventory - BDI*), não são fáceis de serem encontrados. Essa escassez decorre principalmente do processo comumente manual de anotação dos dados, que enfrenta desafios como: (i) alto custo e tempo, especialmente para grandes volumes de dados; (ii) necessidade de especialistas, pois algumas tarefas exigem conhecimento clínico (e.g., médicos rotulando exames); e (iii) subjetividade, quando há divergências entre anotadores.

Large Language Models (LLMs), modelos neurais generativos, normalmente pré-treinados em corpora conforme um determinado idioma, surgem como uma alternativa promissora para apoiar a anotação de dados, pois conseguem gerar saídas estruturadas e detalhadas mesmo com poucos exemplos prévios (Li et al., 2023). Neste contexto, este artigo busca mapear o estado da arte sobre o uso de LLMs na anotação de dados textuais relacionados à depressão ou a transtornos mentais, comparando os principais trabalhos que exploram essa aplicação.

2. Materiais e Métodos

Este estudo é de natureza exploratória e tem foco na identificação de trabalhos primários que contribuíram com datasets anotados no contexto da depressão ou de transtornos mentais. A pesquisa foi realizada nas seguintes bases de dados científicas, considerando os últimos 5 anos: ACM Digital Library, IEEE Xplore, ACL Anthology, Springer, Web of Science, SOL SBC e ScienceDirect. Para isso, a string de busca contou com os termos, em inglês e em português: “Anotação de texto” + “datasets” + “depressão ou transtono mental” + “LLM”. Foram selecionados trabalhos que indicam a fonte dos dados (e.g., Reddit, Twitter/X, notas clínicas), LLMs empregados (como *Generative Pre-trained Transformer* - GPT, *Bidirectional Encoder Representations from Transformers* - BERT ou *Large Language Model Meta AI*- LLama), estratégias de anotação (manual, automatizada ou via *crowdsourcing*) e os rótulos empregados.

3. Resultados e Discussão

A Tabela 1 mostra uma visão comparativa dos trabalhos analisados, descritos a seguir.

O trabalho de Turcan e McKeown (2019) apresenta a criação do dataset Dreaddit, no domínio de estresse, cujos dados foram coletados a partir de 190 mil posts da rede social Reddit, considerando cinco subreddits (abusos, social, ansiedade, PTSD e financeiro). As anotações dos dados foram realizadas via *crowdsourcing* (Amazon MTurk), com um processo de validação feito por dois especialistas humanos. A rotulação final foi definida por votação majoritária para dois rótulos: estresse – sim ou não. O dataset Dreaddit final contém 3.553 instâncias.

Naseem et al. (2022) apresentam o DepSeverity, um dataset rotulado para depressão baseado em uma modelagem ordinal que estabelece níveis hierárquicos para distinguir a severidade do transtorno depressivo em posts do Reddit,

penalizando classificações com discrepâncias severas. O processo de anotação iniciou com dois especialistas clínicos rotulando os posts com base em critérios do BDI e do Manual Diagnóstico Estatístico de Doenças Mentais (DSM-5). Em seguida, LLMs como ALBERT e Longformer foram guiados por prompts estruturados para refinar e automatizar a classificação dos posts com base nos rótulos iniciais e complementar a anotação dos especialistas. Ao final, obteve-se 3.553 registros anotados para quatro classes: ausente, leve, moderada e severa.

Priyadarshana et al. (2023) apresentam o HelaDepDet, um dataset anotado para severidade da depressão. O estudo agregou e balanceou dois grandes corpora públicos - DepSeverity e DepTweet (Kabir et al., 2022) combinando textos e ajustando as proporções de cada nível de severidade para evitar vieses. A principal inovação foi gerar probabilidades (vetores de confiança) em que palavras-chave depressivas, associadas à frequência e à correlação de cada termo, exprimem a severidade do post. Os textos de autorrelato foram anotados com respeito aos quatro níveis descritos anteriormente para formar um dataset com mais de 40.000 instâncias rotuladas.

Gupta et al. (2022) apresentam o dataset PRIMATE, no domínio da triagem de depressão, com dados de 21.000 postagens do subreddit r/depression_help. O processo de anotação reuniu especialistas de saúde mental e anotadores treinados via *crowdsourcing*. Para auxiliar o processo, um modelo BERT foi treinado como classificador binário para identificar respostas ao questionário PHQ-9 no texto. A rotulação consistiu em anotações binárias ("Sim"/"Não") para os critérios do PHQ-9 (e.g., humor deprimido, culpa, sono), alcançando 85% de concordância. O dataset final contém 2.003 instâncias rotuladas para classificar a depressão em 5 níveis de severidade (4 do BDI e moderadamente grave).

Hassan et al. (2024) demonstram a criação do dataset multi-rótulo SPAADE-DR com dados provenientes da fusão de datasets como DepSeverity e Dreaddit. O processo de anotação utilizou LLMs em um cenário zero-shot para gerar rótulos múltiplos, avaliando sistematicamente diversos modelos (e.g., Llama-3, GPT-4o-mini) com três estratégias de prompt distintas (rótulo único, multi-rótulo e irrestrito). A rotulação final foi realizada pela combinação ótima (Llama-3 70b + prompt de rótulo único), resultando em anotações para seis transtornos distintos.

Tabela 1 - Quadro Comparativo sobre Trabalhos com Datasets Anotados para Transtornos Mentais

Trabalho	Fontes de Dados	LLMs Utilizados	Estratégia de Anotação	Classes	Dataset
Turcan e McKeown (2019)	190 mil posts do Reddit	BERT	*Crowdsourcing* + dois especialistas	Estresse: Sim ou Não	Dreaddit
Nassem el al., (2022)	Dataset eRisk	ALBERT e Longformer	Anotação por dois especialistas humanos	Depressão: ausente, leve, moderada ou severa	DepSeverity
Gupta et al. (2022)	21.000 posts do Reddit	GPT-4º	Anotação Sintética com LLMs + três especialistas humanos	9 itens do PHQ-9: Sim ou Não	PRIMATE
Priyadarshana et al. (2023)	DepSeverity + DEPTWEET	BERT, MentalBERT e M-BERT	*Crowdsourcing* + especialistas humanos	Depressão: ausente, leve, moderada ou severa	HelaDepDet
Hassan et al., (2024)	Depseverity-Dreaddit + RMHD	GPT-4o-mini, Llama-3, Mistral NeMo, Phi-3.5-MoE, Gemma-2.	Anotação Sintética com LLMs + especialistas humanos	Depressão, Ansiedade, PTSD, TDAH, transtornos alimentares e ideação suicida	SPAADE-DR

Comparando os trabalhos descritos, observa-se que a criação de datasets anotados para treinamento de modelos em apoio à detecção de sinais transtornos mentais, como a depressão, têm se baseado principalmente na utilização de *crowdsourcing* ou LLMs, mas sempre com validação de especialistas humanos. O Dreaddit e o PRIMATE são exemplos. Outros estudos, como o do DepSeverity e o HelaDepDet, também usaram critérios diagnósticos com base no DSM-5. O papel dos LLMs também se transformou: de um assistente de anotação humana no DepSeverity até a geração sintética de anotações para o estudo de comorbidades no SPAADE-DR.

4. Considerações Finais

O mapeamento realizado neste estudo evidencia o amadurecimento das abordagens que utilizam LLMs na anotação de dados textuais, com pouca ou nenhuma supervisão, mesmo em conjunturas complexas como a depressão. Diante da necessidade de combinar automação e supervisão especializada, a aplicação de LLMs no contexto da depressão pode impulsionar a eficiência e escalabilidade de anotações adequadas. Como trabalho futuro, esta pesquisa se alinha à iniciativa de rotulação do dataset DepreRedditBR (Herculano et al., 2024), que visa anotar um vasto corpus de textos do Reddit em português com base nas quatro classes de severidade da depressão (ausente, leve, moderada ou grave).

5. Agradecimentos

Este trabalho foi desenvolvido com o apoio da FAPESQ-PB.

6. Referências

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR. American Psychiatric Association Publishing, Washington, DC, 5th, text revision edition, 2023.

Gupta, Shrey; Agarwal, Anmol; Gaur, Manas; Roy, Kaushik; Narayanan, Vignesh; Kumaraguru, Ponnurangam; and Sheth, Amit. 2022. Learning to Automate Follow-up Question Generation using Process Knowledge for Depression Triage on Reddit Posts. In Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 137–147, Seattle, USA. Association for Computational Linguistics.

HASSAN, Abdelrahaman A.; HANAFY, Radwa J.; FOUDA, Mohammed E. Automated Multi-Label Annotation for Mental Health Illnesses Using Large Language Models. [S.l.: s.n.], 2024.

HERCULANO, Ayrton Douglas Rodrigues; DE PAULA, Taw-Ham Almeida Balbino; FERNANDES, Damires Yluska de Souza; REGO, Alex Sandro da Cunha. DepreRedditBR: Um conjunto de dados textuais com postagens depressivas no idioma português brasileiro. In: DATASET SHOWCASE WORKSHOP (DSW), 6, 2024, Florianópolis/SC. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 77-90. DOI: https://doi.org/10.5753/dsw.2024.243994.

Hua, Y., Na, H., Li, Z. et al. A scoping review of large language models for generative tasks in mental health care. 2025. npj Digit. Med. 8, 230. Disponível em: https://doi.org/10.1038/s41746-025-01611-4.

Kabir, Md.Mohsinul & Ahmed, Tasnim & Hasan, Bakhtiar & Laskar, Md Tahmid Rahman & Joarder, Tarun & Mahmud, Hasan & Hasan, Md Kamrul. 2023. DEPTWEET: A Typology for Social Media Texts to Detect Depression Severities. Computers in Human Behavior, Volume 139, 107503, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2022.107503.

Li, Minzhi; Shi, Taiwei; Ziems, Caleb; Kan, Min-Yen; Chen, Nancy F.; Liu, Zhengyuan; Yang, Diyi. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. 2023. Disponível em: https://aclanthology.org/2023.emnlp-main.92.pdf.

Priyadarsana, Y. P.; Liang, Z.; Piumarta, I.. HelaDepDet: A Novel Multi-class Classification Model for Detecting the Severity of Human Depression. Collaboration Technologies and Social Computing. CollabTech 2023. Lecture Notes in Computer Science, vol 14199. Springer, Cham. https://doi.org/10.1007/978-3-031-42141-9_1

RANI, Saima; AHMED, Khandakar; SUBRAMANI, Sudha. From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives. Applied Sciences, v. 14, n. 1547, 2024. Disponível em: https://www.mdpi.com/journal/applsci. Acesso em: 18 mai. 2025.

Seo, M., Baek, J., Thorne, J., & Hwang, S. J. (2024). Retrieval-augmented data augmentation for low-resource domain tasks. *arXiv preprint arXiv:2402.13482*.

Turcan, Elsbeth and McKeown, Kathleen. Dreaddit: A reddit dataset for stress analysis in social media. arXiv preprint arXiv:1911.00133, 2019.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. Early identification of depression severity levels on reddit using ordinal classification. In Proceedings of the ACM Web Conference 2022, pages 2563–2572, 2022.