

DepressSet: Um conjunto de dados de análises textuais sobre postagens depressivas

Silas Lima Filho, Eliel Roger da Silva,
Jonice Oliveira, Mônica Ferreira da Silva

¹Instituto de Computação (IC)
Programa de Pós-Graduação em Informática (PPGI),
Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brasil

silaslfilho@ppgi.ufrj.br, elielsilva@ufrj.br

monicaferreiramail@gmail.com, jonice@dcc.ufrj.br

Abstract. *Social media can be useful for seeking help or guidance on how to deal with or to better understand depressive disorder. However, dealing with data on such a disorder can be challenging due to the sensitivity of the content domain or even the difficulty in finding data on the subject. In this work, we present a dataset collected from depression-related communities on Facebook in September 2022. We specify the extraction, processing, storage, and release of the data, along with its limitations, challenges, and lessons learned. We enriched the captured data with linguistic analyses of the posts and also with predictions for each post using a text classification model. Finally, we propose potential applications of the dataset and its limitations.*

Resumo. *Mídias sociais podem ser úteis para buscar ajuda, ou orientação sobre como lidar, ou para entender melhor o transtorno depressivo. No entanto, lidar com dados sobre tal transtorno pode ser um desafio pela sensibilidade do domínio do conteúdo, ou mesmo pela dificuldade de encontrar dados a respeito do tema. Neste trabalho apresentamos um conjunto de dados coletados de comunidades sobre depressão no Facebook no mês de Setembro de 2022. Especificamos a extração, tratamento, armazenamento e abertura dos dados, com suas limitações, desafios e aprendizados. Enriquecemos os dados capturados com análises linguísticas das postagens, e também com a predição de cada postagem utilizando um modelo de classificação textual. Finalmente, encaminhamos propostas de potenciais aplicações do conjunto de dados e suas limitações.*

1. Introdução

As mídias digitais, sendo utilizadas como ferramentas de comunicação entre diferentes grupos da sociedade, se tornam cada vez mais fontes de informação e conscientização sobre problemas e preocupações [De Choudhury 2014]. Usuários podem buscar ajuda sobre doenças em plataformas online, como mídias sociais, fóruns de apoio à saúde, *websites* de organizações, entre outros. Ainda nesse contexto temos os transtornos mentais, que afetam uma parcela significativa da população. Por exemplo, de acordo com a Organização Mundial de Saúde (OMS), cerca de 300 milhões de pessoas de diferentes idades sofrem de algum nível de depressão¹.

¹www.who.int/en/news-room/fact-sheets/detail/mental-disorders [acesso 08-01-2024]

Apesar do potencial das mídias sociais para análise de transtornos mentais, a obtenção de dados para pesquisas nessa área é um desafio [Ríssola et al. 2020]. A coleta e rotulação de grandes quantidades de dados de mídias sociais são processos complexos e demorados.

Apresentamos as etapas de construção e o artefato *DepressSet*², um conjunto de dados de postagens de comunidades sobre depressão no Facebook, plataforma escolhida por sua popularidade [Chou et al. 2021]. O dataset inclui metadados das mensagens (curtidas, reações, etc.), análises linguísticas e classificação por um modelo treinado.

Aplicamos enriquecimento semântico ao dataset para garantir a proveniência dos dados, usando a biblioteca *Prov*³, que implementa o modelo W3C Prov-DM⁴, uma abordagem similar ao de [Gonçalves et al. 2021]. O processo de construção do dataset e o grafo de proveniência estão disponíveis no repositório. Para atender às diretrizes FAIR, o dataset está no Zenodo, identificado pelo DOI 10.5281/zenodo.11209982, e descrito em um dicionário de dados.

A Seção 2 apresenta trabalhos relacionados; a Seção 3 apresenta o método, seu protocolo e etapas relacionadas para construção do dataset; a seção 4 apresenta uma breve discussão sobre os aspectos éticos; a Seção 5 apresenta limitações, discussões, encaminhamentos, e considerações finais.

2. Trabalhos Relacionados

Não é atual o esforço em tentar usar métodos computacionais para abordar o transtorno depressivo. Não obstante, não se perdeu o valor e a importância de tal tarefa. Estudos que abordam a identificação de usuários depressivos nas mídias sociais utilizam algoritmos e técnicas de reconhecimento de padrões, aproveitando o Processamento de Linguagem Natural (PLN) para realizar uma análise sistêmica de textos em postagens de mídias sociais. Eles se aprofundam em aspectos técnicos da pragmática computacional, incorporando informações e conhecimentos da área da Saúde, como Psicologia e Medicina, e utilizam questionários psicométricos [Giuntini et al. 2020]. O trabalho de [P. Lima Filho et al. 2022] por exemplo explora features usadas em modelos de classificação com profissionais da área da saúde.

Já outros autores, lidaram com a questão através da criação de conjunto de dados especializado com dados provenientes de mídias sociais. [Shing et al. 2018] criaram um conjunto de dados para estudar a avaliação do risco de suicídio no Reddit. Postagens entendidas com conteúdo de suicídio, ou seja, usuários em risco de suicídio, foram coletadas com base em sua participação em um fórum de discussão chamado *Suicide Watch*. Por fim, [Ríssola et al. 2020] apresentam uma metodologia para coletar automaticamente um grande conjunto de dados de publicações sobre depressão e não depressão de plataformas de mídia social.

O presente trabalho se diferencia das demais abordagens nos seguintes aspectos: Uso do Facebook⁵ como fonte de dados, utilizando dessa forma uma das mídias

²Dataset Zenodo

³<https://prov.readthedocs.io/en/latest/readme.html>

⁴www.w3.org/TR/prov-dm/

⁵<https://facebook.com>

sociais com maior aderência entre usuários nos EUA [Chou et al. 2021]; Uso de metadados e valores classificando as publicações de acordo com o conteúdo; Conformidade com os princípios FAIR, visando garantir a reprodutibilidade da construção do conjunto de dados.

3. Método de Construção do Dataset

Descrevemos abaixo as etapas para a construção do dataset. As etapas são: Captura de Dados na seção 3.1, Features linguísticas na seção 3.2, e *predição de classes* na 3.3. A Figura 1 simplifica e resume as etapas e atividades do método de construção do dataset. A construção do presente dataset faz parte de um projeto maior de pesquisa que usa a abordagem de *Design Science Research* como epistemologia do método científico, e melhor explicado e detalhado em [Lima Filho. et al. 2024].

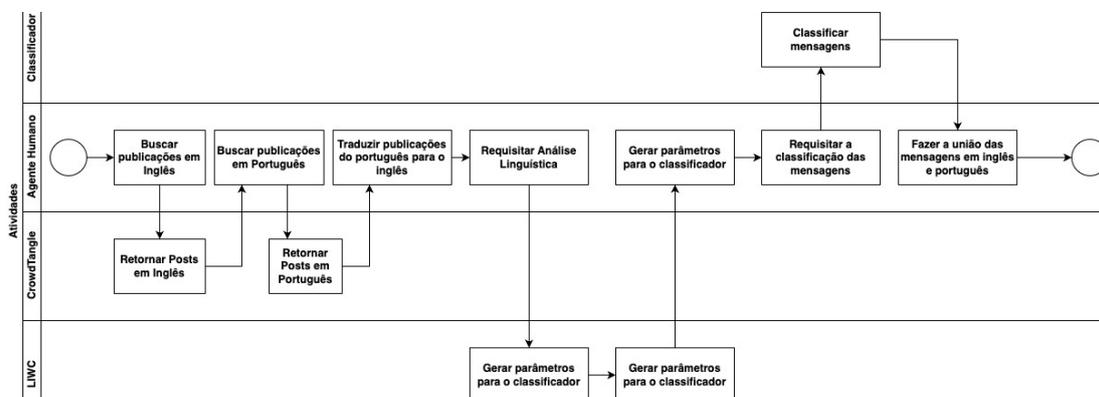


Figura 1. Fluxograma de construção do dataset.

3.1. Captura de Dados

Para construir o dataset, foram explorados textos capturados de mídias sociais usando a ferramenta Crowdtangle, utilizada para pesquisa e coleta de dados de perfis e grupos públicos⁶. Também é usada para verificar links e monitorar postagens. Os dados foram coletados de comunidades abertas do Facebook que discutiam diretamente sobre depressão ou tópicos relacionados em inglês e português, durante setembro de 2022, um período em que ações de mídia geralmente promovem o cuidado com a saúde mental. A busca foi limitada a grupos abertos (comunidades) e postagens do tipo *status*, excluindo publicações pagas, perfis verificados e páginas do Facebook. Inicialmente, termos relacionados à depressão foram definidos com base na literatura. Os termos em português e inglês foram:

- “*quero morrer*”, *depressão*, *deprimid*, *depressiv*, *angustia*
- *depression*, *depressed*, *depressive*, *anguish*, *distress*

Buscou-se por publicações de status em comunidades que discutem diretamente sobre depressão para obter refinamento nas publicações retornadas. Foram coletadas 835 publicações de 4 comunidades em inglês e 1945 publicações de 7 comunidades em português (4 sobre depressão, 2 sobre tratamento psicológico e 1 sobre

⁶www.crowdtangle.com/resources [acessado em 11-03-2024]

luto). Após essa etapa, os dados coletados foram explorados e os atributos relevantes selecionados como categóricos ou contínuos. A Tabela 3.1 descreve o dicionário de dados com os atributos do dataset até essa etapa, onde o atributo categórico que identifica o autor do post é nulo, preservando dados sensíveis.

Categoria	Atributo	Descrição	Tipo do Dado
Contínuo	Total Interactions	Qtd total de interações da postagem	Núm. Inteiro
	Likes	Qtd total de curtidas da postagem	Núm. Inteiro
	Comments	Qtd total de comentários da postagem	Núm. Inteiro
	Shares	Qtd total de compartilhamentos da postagem	Núm. Inteiro
	Love	Qtd total de reações amei da postagem	Núm. Inteiro
	Wow	Qtd total de reações surpresa da postagem	Núm. Inteiro
	Haha	Qtd total de reações risos da postagem	Núm. Inteiro
	Sad	Qtd total de reações triste da postagem	Núm. Inteiro
	Angry	Qtd total de reações raiva da postagem	Núm. Inteiro
	Care	Qtd total de reações cuidado da postagem	Núm. Inteiro
	Post Views	Qtd de visualizações da postagem	Núm. Inteiro
Categórico	Message	Mensagem da postagem	Texto
	Group Name	Nome do grupo onde a postagem foi feita	Texto
	Facebook Id	Identificador da postagem	Texto
	Post Created	Data de criação da postagem	Data e hora
	URL	Link da postagem	Texto link

Tabela 1. Dicionário de dados com os atributos obtidos na etapa de coleta.

3.2. Features Linguísticas

Uma parte dos atributos do dataset são resultados de análises linguísticas, feitas com o léxico *Linguistic Inquiry and Word Count* (LIWC), que processa e analisa textos usando um léxico predefinido. Foi utilizado apenas o atributo “*Message*”, e as mensagens foram traduzidas do português para o inglês. Aplicou-se a versão mais recente (2022) do dicionário léxico em ambas as comunidades. A análise textual usou o módulo principal da ferramenta, quantificando quatro aspectos: Pensamento Analítico (“*Analytical Thinking*”) que captura o grau de uso de palavras que sugerem padrões formais e lógicos; Influência (“*Clout*”), que se refere ao status social, confiança ou liderança demonstrados na escrita; Autenticidade (“*Authenticity*”), que expressa espontaneidade no discurso; e Tom Emocional (“*Emotional Tone*”), onde scores altos indicam um tom positivo e scores baixos, um tom negativo. A Tabela 2 descreve os atributos do dicionário de dados obtidos com a análise textual do LIWC.

Categoria	Atributo	Descrição	Tipo do Dado
Contínuo	Segment	Qtd total de interações da postagem	Núm. Inteiro
	WC	Qtd total de palavras da postagem	Núm. Inteiro
	Analytic	Métrica pensamento analítico, lógico	Núm. Real
	Clout	Métrica que mede status, liderança	Núm. Real
	Authentic	Métrica que mensura honestidade e genuinidade	Núm. Real
	Tone	Grau de tom positivo ou negativo da postagem	Núm. Real
	WPS	Média de palavras por sentença	Núm. Real
	BigWords	Percentual de palavras com, ou mais de 7 letras	Núm. Real

Tabela 2. Dicionário de dados obtidos na etapa de análise linguística.

3.3. Classificação de postagens

Para enriquecer o conjunto com a classificação se a postagem indica sinal de transtorno depressivo ou não, desenvolvemos um modelo de classificação treinado sobre os

atributos linguísticos obtidos na seção 3.2. Nosso modelo de classificação é treinado com dados de uma fonte cientificamente fundamentada [Low et al. 2020], usando como base o modelo proposto por [Yen et al. 2021], que utiliza os sentimentos de uma postagem em mídia social como pesos em um processo de aprendizado ativo, onde os pesos são usados como reforços. Usamos postagens de três subfóruns distintos: *depression*, *fitness* e *divorce*, rotuladas como não depressivas (0), enquanto que as postagens do subfórum de depressão foram rotuladas como positivas (1). Incluímos postagens de fóruns que tinham potencial de discussões relacionadas à depressão, o que apresenta um desafio na distinção eficaz entre textos positivos e negativos. Para cada postagem no dataset, obtemos as métricas de *Emotional Tone*, *Analytical Thinking*, *Clout* e *Authentic*. Utilizamos o modelo de linguagem BERT para vetorizar as postagens, usando a biblioteca *sentence-transformers* [Reimers e Gurevych 2019]. Ressalta-se que os dados classificados não possuem um rótulo prévio que asseste se a postagem é feita por alguém diagnosticado com o transtorno depressivo. O valor do atributo predição para cada instância do dataset é gerado com o classificador treinado em outro momento da pesquisa. A Tabela 3 apresenta o atributo “predição” que foi inserido no dataset.

Categoria	Atributo	Descrição	Tipo do Dado
Catagórico	Predição	Classificação se a postagem indica sinais de depressão	Binário

Tabela 3. Dicionário dos dados obtidos na etapa de classificação.

A Tabela 4 apresenta a quantidade de postagens de cada comunidade e também a quantidade de reações total das postagens de cada comunidade. Exibindo assim as comunidades que contém maior participação e engajamento de seus usuários.

Tabela 4. Comparação dos Grupos de Suporte

Comunidade	Qtd de Postagens	Qtd Interações
A Dor da perda	66	3655
Anxiety & Depression Quotes	69	2616
DEPRESSÃO & ANSIEDADE NÃO É FRESCURA	238	13992
DEPRESSÃO E ANSIEDADE 🙁💔	1562	68432
Depression and Anxiety	214	10774
Depressão, Ansiedade e Eu	13	672
Mental Health Awareness and Support Group	186	8719
Suicidal Thoughts	366	16923
TCC - TERAPIA COGNITIVO COMPORTAMENTAL (PROFISSIONAIS ENGAJADOS)	16	874
TCC - Terapia Cognitivo-Comportamental	29	2395
depressão 🙁😭	20	1431

4. Aspectos éticos

Para manter a privacidade dos autores das mensagens e a sensibilidade dos dados, optamos por omitir os nomes dos usuários e o conteúdo textual das mensagens. Portanto, nosso dataset não identifica os autores, focando apenas nos resultados das análises linguísticas, na classificação das mensagens e nos dados das reações às postagens. Para uso dos dados em pesquisa, o acesso ao dataset é disponibilizado apenas após um processo de verificação e acordo de ciência do requerente, responsabilizando-o a não identificar os autores das postagens. Durante a busca

dos dados, tivemos a autorização do comitê de ética de pesquisa de nossa instituição, pois para construir um modelo de classificação, era necessário interagir com dados e participantes humanos. O CAAE da submissão na Plataforma Brasil é 54865821.5.0000.5263, e o número do parecer é 5.225.791.

5. Limitação, discussão e considerações finais

Concluindo assim esse artigo, disponibilizamos online e abertamente o dataset contendo 2780 registros com os resultados de análises linguísticas de postagens relacionadas ao fenômeno depressão de comunidades do Facebook ⁷, respeitando os princípios de ciência aberta. Os dados foram coletados durante o período de Setembro de 2022, de comunidades de falantes dos idiomas português e inglês. Os dados foram enriquecidos com o resultado de análises linguísticas obtidas com o LIWC, e também classificados indicadores ou não de sintomas de depressão. Também enriquecemos nosso processo de construção nomeando metadados sobre todo o processo, atividades e agentes, dando assim proveniência aos dados. Além dos dados, disponibilizamos o grafo de proveniência do dataset onde descrevemos as tarefas, agentes e versões do dataset durante o processo até chegar na versão final do conjunto de dados.

Por mais que os dados sejam limitados a um determinado período de ano, o dataset permite a exploração de outros pesquisadores dos dados resultantes das análises textuais. Podendo assim serem explorados ou combinados com outras informações, ou então utilizados no treinamento de modelos de classificação ou regressão dos dados aqui apresentados. Outras aplicações que podem ser desenvolvidas, é a verificação e validação desses dados com outros domínios, ou então dados do mesmo cenário mas provenientes de outras mídias sociais. Recomendamos sempre que a utilização de dados desse domínio seja feita de modo responsável e ético por parte dos pesquisadores por se tratar de dados relacionados à saúde mental de pessoas.

6. Agradecimentos

O presente trabalho foi apoiado pela CAPES Edital nº 09/2020 - Proc. nº 223038.014313/2020-19, e parcialmente apoiado pelo programa *Oracle for Research* (nº prêmio CPQ-2160239).

Referências

- [Chou et al. 2021] Chou, W.-Y. S., Gaysynsky, A., Trivedi, N., e Vanderpool, R. C. (2021). Using social media for health: National data from hints 2019. *Journal of Health Communication*, 26(3):184193.
- [De Choudhury 2014] De Choudhury, M. (2014). Opportunities of social media in health and well-being. *XRDS*, 21(2):2327.
- [Giuntini et al. 2020] Giuntini, F. T., Cazzolato, M. T., de Jesus Dutra dos Reis, M., Campbell, A. T., Traina, A. J. M., e Ueyama, J. (2020). A review on recognizing depression in social networks: challenges and opportunities. *Journal of Ambient Intelligence and Humanized Computing*, 11:1–17.

⁷Dataset Zenodo[acesso 25-03-2024]

- [Gonçalves et al. 2021] Gonçalves, M. V., dos Santos, J., Ferreira, C., Zavaleta, J., Cruz, S., e Oliveira, J. (2021). Datasets curados e enriquecidos com proveniência da campanha nacional de vacinação contra covid-19. pp. 148–159.
- [Lima Filho. et al. 2024] Lima Filho, S., Ferreira da Silva, M., e Oliveira, J. (2024). A systematic analysis of depression-related discourse within facebook: A comparison between brazilian and american communities. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF*, pp. 466–473. INSTICC, SciTePress.
- [Low et al. 2020] Low, D. M., Rumker, L., Torous, J., Cecchi, G., Ghosh, S. S., e Talkar, T. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.
- [P. Lima Filho et al. 2022] P. Lima Filho, S., Ferreira da Silva, M., Oliveira, J., e Ruback, L. (2022). A study about gathering features in depression detection problem with health professionals community. *iSys - Brazilian Journal of Information Systems*, 15(1):10:110:26.
- [Reimers e Gurevych 2019] Reimers, N. e Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- [Ríssola et al. 2020] Ríssola, E. A., Bahrainian, S. A., e Crestani, F. (2020). A dataset for research on depression in social media. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*.
- [Shing et al. 2018] Shing, H.-C., Nair, S., Ziriky, A., Friedenber, M., Daumé III, H., e Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. In Loveys, K., Niederhoffer, K., Prud’hommeaux, E., Resnik, R., e Resnik, P., editors, *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 25–36, New Orleans, LA. Association for Computational Linguistics.
- [Yen et al. 2021] Yen, S.-C., Chu, K.-C., e Tsai, P.-Y. (2021). Prediction model of social network suicide ideation by small sample. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 385–389. IEEE.