BrStudentMH: A Reddit Dataset for Mental Health Analysis in the Brazilian University Context

Maicon C. Marques[1], **Camila O. Silva**[1], **Leonardo P. Dallagnol**[1], **Tiago C. B. Rocha**[1], **Jean R. Ponciano**[1]

[1] Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
São Carlos – SP – Brazil

{maiconchavesmarques, caamilaoliv, dallagnol_leonardo, tiagocbr}@usp.br,
jeanponciano@icmc.usp.br

Abstract. *The increasing use of social networks has made platforms like* REDDIT *spaces for discussing sensitive topics without fear of judgment. This paper presents* `BrStudentMH`*, a dataset comprising 943 mental health-related posts and 15,680 comments in Brazilian Portuguese, collected from Brazilian higher education-related subreddits. The data were obtained through the* REDDIT API *and processed in a pipeline consisting of collection, categorization using Large Language Models, and pseudonymization. Its application is expected across various fields, providing insights into the mental health of Brazilian university students and contributing to the development of support and treatment policies.*

1. Introduction

The preservation of mental health is increasingly recognized as a foundational pillar for human development and global well-being, particularly during the transition into adulthood [Guthold et al. 2023]. Within the academic environment, the psychological pressure associated with rigorous performance expectations, social competition, and the pursuit of professional certification has led to a significant rise in mental health issues among university students [Li et al. 2025]. Research indicates that these students face a higher risk of developing mental health challenges — specifically depression, anxiety, and stress —— when compared to their non-student peers [Voss et al. 2023]. In this high-pressure context, positive social connectivity significantly helps university students mitigate these psychological burdens and improves their overall academic persistence [Ruihua et al. 2025].

To find this necessary support, many university students have turned to digital spaces like REDDIT, which functions as a social tool where individuals can share experiences and seek advice within specialized communities. Considering the Brazilian context, students frequently utilize REDDIT to share personal experiences, exchange advice, and pose questions regarding the specific challenges of the academic environment. The anonymity provided by platforms like this fosters a greater sense of confidence among users, encouraging them to disclose intimate opinions, personal experiences, and sensitive advice that they might otherwise withhold in non-anonymous settings [Croes et al. 2024].

However, despite the richness of this spontaneous data for sentiment analysis and mental health research, there is a notable absence of organized, publicly accessible datasets that focus on the Brazilian academic environment. Most existing resources are either English-centric or lack the domain-specific labels required to categorize academic-related mental health discourse.

To address this scarcity, this paper presents `BrStudentMH`, a novel dataset curated to reflect the discourse on mental health within the Brazilian academic sphere. The construction of this resource followed a streamlined multi-stage pipeline: (i) extraction, where the REDDIT API was used to collect data from a targeted selection of academic-related Brazilian subreddits; (ii) categorization, involving Large Language Models (LLMs) to classify and validate the contents of the posts; and (iii) pseudonymization, where sensitive data is removed to ensure the privacy of each user on the platform. The resulting dataset comprises 943 posts and 15,680 comments in Brazilian Portuguese (PT-BR), featuring a relational identification system that pairs each post with its respective comments to ensure full traceability and organization.

2. Related works

Due to the growing interest in mental health analysis, several studies have been proposed to support research in this domain. In this context, the availability of structured datasets on mental health conditions has become an important foundation for enabling and advancing future investigations. For example, the `BrSuicides` dataset [Klann and Sousa 2024] provides a dataset derived from DATASUS records, which underwent cleaning, preprocessing, and enrichment procedures to support analyses of suicide cases in Brazil.

As a source of mental health-related information, social media platforms stand out due to the amount of user-generated content, such as posts, comments, and discussions. The study of [Coppersmith et al. 2015] exemplifies how social media content can be used to construct mental health datasets. The authors extend previous work by extracting data based on self-reported diagnoses from TWITTER (now called X) and compiling a large-scale dataset encompassing a wide range of conditions.

In the Brazilian Portuguese context, `SetembroBR` [Santos et al. 2023] compiles data from TWITTER into two distinct datasets to support the development of depression and anxiety detection models. Similarly, data from the YAHOO! ANSWERS platform were used to construct the `Depress-pt-br` dataset [da Silva Nascimento et al. 2018]. This dataset consists of textual interactions categorized into depressive and non-depressive instances, aiming to identify linguistic patterns and behavioral signals associated with depression. Despite the historical value, platforms such as YAHOO! ANSWERS have been discontinued, and the short-form nature of tweets limits the depth of content analysis.

To capture longer, more introspective, and anonymized user discourses, researchers have increasingly turned to REDDIT. The `SMHD (Self-Reported Mental Health Diagnoses)` dataset, for instance, was introduced using REDDIT posts from individuals who self-report one or more mental health conditions, paired with a control group of users with no evidence of mental health-related content [Cohan et al. 2018]. However, the focus is on exploring online language usage across multiple mental health conditions in English. Furthermore, `Dreaddit` was presented as a text corpus for the identification of stress in social media content, consisting of posts from different categories of REDDIT communities [Turcan and McKeown 2019].

Further exploring the English-language context on REDDIT, the `Reddit Mental Health Dataset` was proposed to evaluate the impact of the COVID-19 pandemic across various subreddits using natural language processing and topic modeling techniques [Low et al. 2020]. Similarly, a distinct dataset was constructed from men-

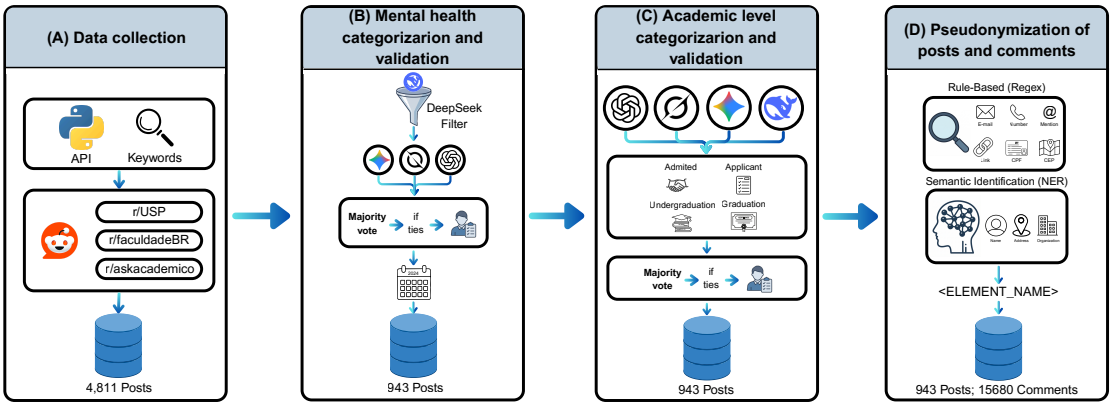

Figure 1. Proposed pipeline. (A) Data collection. (B) Mental health categorization and validation. (C) Academic level categorization and validation. (D) Pseudonymization of posts and comments.

tal health-focused subreddits by annotating posts into severity levels (absent, moderate, or severe) to train machine learning classifiers [Sampath and Durairaj 2022]. While these datasets demonstrate the effectiveness of REDDIT for text mining, their restriction to the English language and to general audiences limits their use in more specific contexts.

To address the linguistic gap and leverage the advantages of forum-based discussions in Brazilian Portuguese, recent efforts have begun to explore REDDIT communities in Brazil, although the number of available datasets remains limited. For instance, `DepreRedditBR` was introduced as a dataset of depressive posts in Portuguese collected from REDDIT [Herculano et al. 2024]. Unlike approaches that rely solely on self-reported diagnoses, this dataset gathers texts based on medical terms validated by domain experts. Despite representing a significant advancement, `DepreRedditBR` targets a general audience and does not focus on specific groups, leaving a critical gap for datasets tailored to academic environments in Brazil.

Motivated by this gap, this work introduces a novel dataset specifically designed to capture the mental health discourse of Brazilian university students. By focusing on academic communities within REDDIT, the proposed dataset can be used, for example, in the investigation of linguistic patterns and behavioral signals in a long-form, more targeted and context-specific setting.

3. Methodology

Figure 1 outlines the proposed pipeline for creating the `BrStudentMH` dataset. We begin by collecting posts and comments from three Brazilian higher-education subreddits using specific mental health-related keywords (Fig. 1(A)). Subsequently, LLMs are employed to refine the dataset by excluding irrelevant posts (Fig. 1(B)) and inferring the academic levels reflected in the post (Fig. 1(C)). Finally, posts and comments undergo pseudonymization, ensuring the safe disclosure of information (Fig. 1(D)). Further details are provided below.

3.1. Dataset Construction

The dataset targeted three higher education-related Brazilian subreddits, chosen due to their complementary nature in terms of academic journey and diversity coverage: *r/USP*

(approximately 65 thousand members), *r/faculdadeBR* (approximately 84 thousand members), and *r/askacademico* (approximately 13 thousand members). The *r/USP* subreddit, dedicated to the University of São Paulo (USP), one of the country's largest universities and frequently ranked among the best in Latin America in international rankings [Times Higher Education 2025], contains posts more directed toward topics related to entrance exams and undergraduate studies, usually associated with admission and life at the university. The *r/faculdadeBR* subreddit presents a more general perspective, addressing topics from several different universities, both public and private, and from different regions of Brazil. Finally, *r/askacademico* frequently covers topics related to graduate studies, doctorates, specializations, and more advanced stages of academia.

Inspired by similar keyword-based searches [Low et al. 2020], we derived a set of 65 distinct mental health-related keywords, which are available in the dataset's GitHub repository. Using a Python script, the PRAW[1] library (The Python Reddit API Wrapper) was then used to collect posts that contained at least one of the defined keywords, along with several of their metadata, including comments, resulting in a total of 4,811 posts retrieved. Fig. 1(A) summarizes the collection process.

3.2. Data Validation and Enrichment with LLMs

As many of the retrieved posts were not directly related to mental health, we performed an initial refinement using DeepSeek V3 to categorize each post as related or not to mental health. This model was chosen due its strong performance on similar tasks [Xu et al. 2025], and the prompt included a task description, an explanation of what the posts referred to, the possible categories, and examples for each category. After this filtering step, the number of posts was reduced to 1,141.

To further validate the relevance of the resulting posts to our dataset, three additional models — ChatGPT 4o, Grok 3, and Gemini Pro 2.5 — independently performed the same categorization task on this new set of posts. The convergence rate was 81%, meaning that all models produced the same categorization in 81% of the posts. The final categorization was defined through majority voting, and human evaluation was conducted to solve ties. From the 1,070 retained posts, we removed those dated before 2024 due to their temporal sparsity and small quantity. Fig. 1(B) summarizes this mental health categorization and filtering step. The final dataset contains 943 posts and 15,680 comments. The posts cover 476 days, from Jan. 1, 2024, to April 20, 2025, while the associated comments span a broader time window, ranging from Jan. 1, 2024, to Sep. 15, 2025.

To deepen data understanding and enable new analyses, a second categorization step was conducted to infer the academic level from the post's title and content. Four possible levels were defined to cover the complete journey within the university context. The *applicant* academic level refers to posts related to admission exams for higher education, frequently involving entrance examinations such as the *Exame Nacional do Ensino Médio (ENEM)* or exams of *Fundação Universitária para o Vestibular (FUVEST)*. The *admitted* level focuses on posts from students who had already taken the application and were accepted but had not yet started classes. The *undergraduate* level considers posts involving the daily routine of students at this level of study, including exams, course schedules,

[1] `https://praw.readthedocs.io/en/stable/`

course failures, and subjects. The *graduate* level encompasses more advanced academic study, such as master's and doctoral degrees, specializations, and related programs.

The academic level categorization was independently performed by the same four models as before, with the final category defined by majority voting (or human judgment in case of ties). The prompt was also similar to the previous one, defining the four possible levels and examples of situations that characterized each one. In 68% of the publications, all models assigned the same category. Fig. 1(C) summarizes this step. The performance of the models across these two categorization steps will be discussed in Section 7.

3.3. Data Pseudonymization

To pseudonymize sensitive information, a combination of two identification techniques was used, following approaches previously tested in the literature [An et al. 2025]. The first consists of regular expressions (Regex), which enable the identification of textual patterns with well-defined characteristics. The second is the employment of the Named Entity Recognition (NER) model *ner-bert-large-cased-pt-lenerbr* [Guillou 2021], which addresses patterns that are not easily identifiable through rigid rules, such as names of people, places, and organizations.

In the first stage, regular expressions were used to find, in the posts (title and body) and in the comments (body), elements such as personal email addresses, phone numbers, URLs, mentions of other users, and references to specific social media profiles. Values that could lead to personal identification were replaced with a tag `<ELEMENT_NAME>` (e.g., `<PHONE>`, `<URL>`, `<MENTION>`); the others were not replaced.

Using Regex, we also searched for the numbers of Brazilian personal documents such as *Cadastro de Pessoas Físicas (CPF), Registro Geral (RG), Cadastro Nacional da Pessoa Jurídica (CNPJ), Carteira Nacional de Habilitação (CNH)*, and passports, as well as zip codes *(Códigos de Endereçamento Postal (CEPs))* and Internet Protocol (IP) addresses, which were replaced with their respective tags following the same procedure.

In the second stage, we used NER to cover aspects not addressed by Regex. Three types of sensitive information were considered: personal names, personal addresses, and organization names. After analyzing the obtained results, no instances of personal addresses were identified, nor any mention of organizations that could be considered harmful. Based on this, it was decided to mask only personal names according to the NER classification. Therefore, personal names were replaced with `<PERSON>` in the dataset.

Beyond the elements described above, any information that could facilitate the identification of specific users was also masked. Post IDs, comment IDs, and usernames were replaced to preserve pseudonymity. Despite these substitutions, the relationships among posts, comments, and their replies remain intact, as only the identifiers were modified, not removed. Furthermore, URLs in posts and comments were eliminated, and publication timestamps were reduced to the date only, with hours and seconds discarded. Fig. 1(D) summarizes this pseudonymization step.

4. Data Organization

Tables 1 and 2 depict the dataset elements, showing the types and brief descriptions of each field for posts and comments, respectively. The files were made available in the

".json" format, in which each field acts as a key that returns the corresponding associated information. Additionally, a simplified entity–relationship diagram can be observed in Fig. 2, highlighting the relationships between the two files and their corresponding fields.

The dataset is available at `https://github.com/MaiconChavesMarques/BrStudentMH-dataset`, with both keywords and academic levels written in PT-BR, as already occurs with the post title and body, and the comment body. The values for the academic levels are: `VESTIBULAR`, `INICIO`, `GRADUAÇÃO`, and `POS-GRADUAÇÃO` (Applicant, Admitted, Undergraduate, and Graduate, respectively).

Table 1. Data dictionary for the Posts

Field	Type	Description
id	string	Alphanumeric identifier that uniquely identifies a post.
title	string	Title of the post created by the user, summarizing the main content or topic.
body	string	Textual content of the post, which may include reports, questions, rants, etc.
date	date	Post publication date in YYYY-MM-DD format.
username	string	Unique identifier of the user who created the post.
score	int	Post score based on user interactions (e.g., upvotes and downvotes).
academic_level	string	Academic level the post refers to (e.g., Applicant, Admitted, Undergraduate, Graduate).
subreddit	string	Name of the subreddit to which the post belongs.
keyword	string	Keyword used in the search that retrieved the post.

Table 2. Data dictionary for the Comments

Field	Type	Description
id	string	Alphanumeric identifier that uniquely identifies a comment.
post_id	string	Identifier of the post to which the comment belongs.
parent_id	string	Identifier of the comment being replied to. If null, it indicates a direct reply to the post.
username	string	Unique identifier of the user who made the comment.
date	date	Comment publication date in YYYY-MM-DD format.
body	string	Textual content of the comment, which may include reports, questions, rants, etc.
score	int	Comment score based on user interactions (e.g., upvotes and downvotes).

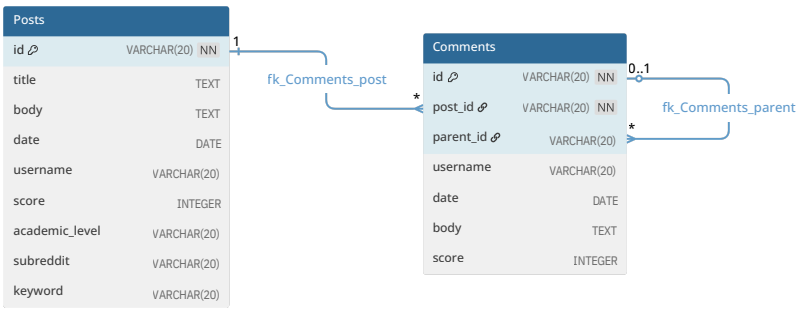

Figure 2. Entity Relationship Diagram

5. Descriptive and Exploratory Analysis

This section presents a descriptive and exploratory analysis of the dataset.

5.1. Posts

The dataset comprises **943** posts distributed across three subreddits, 52 keywords, 841 unique users, and four inferred academic levels (Tab. 3). The subreddit *r/faculdadeBR* concentrates the largest number of posts (478), accounting for more than half of the

Table 3. Number of posts by subreddit, unique users, and academic level

Category	`r/askacademico` (Total / Unique)	`r/faculdadeBR` (Total / Unique)	`r/USP` (Total / Unique)	**Overall** (Total / Unique)
`Applicant`	4 / 4	94 / 89	159 / 145	257 / 234
`Admitted`	0 / 0	60 / 54	69 / 65	129 / 119
`Undergraduate`	34 / 32	310 / 288	130 / 118	474 / 434
`Graduate`	59 / 55	14 / 14	10 / 10	83 / 79
Overall	**97 / 89**	**478 / 438**	**368 / 323**	**943 / 841**

Table 4. Top 5 keywords by number of posts and unique users

Keyword	Posts (%)	Unique Users
`Fear` (*medo*, in PT-BR)	210 (22.27%)	197
`Anxiety` (*ansiedade*, in PT-BR)	149 (15.80%)	140
`Sadness` (*tristeza*, in PT-BR)	62 (6.57%)	58
`Doubt` (*dúvida*, in PT-BR)	42 (4.45%)	42
`Depression` (*depressão*, in PT-BR)	39 (4.14%)	36

dataset. A similar concentration is observed for the academic level, with 474 posts labeled as *undergraduate*. Keyword usage is also uneven, with higher frequencies for terms such as *fear* and *anxiety* (Tab. 4).

Regarding textual content, post length varies across categories, reflecting different discussion patterns (Tab. 5). Categories such as *r/askacademico* and *graduate* present longer texts (over 1200 characters on average), while *admitted* and *r/USP* tend to have shorter posts. In contrast, title length remains stable across categories, ranging from 38 to 51 characters (approximately five to eight words).

To complement the quantitative analysis, Figure 3 shows word clouds of post titles and bodies (in PT-BR). They reinforce previously observed patterns, with frequent terms such as fear (*medo*, in PT-BR), doubt (*dúvida*), and vent (*desabafar*), alongside academic-

Table 5. Mean and median length of posts and titles by category

Category	**Body (Chars)** (Mean / Median)	**Body (Words)** (Mean / Median)	**Title (Chars)** (Mean / Median)	**Title (Words)** (Mean / Median)
`r/askacademico`	1215 / 1143	208 / 190	50 / 44	8 / 7
`r/faculdadeBR`	1090 / 861	190 / 150	45 / 38	8 / 7
`r/USP`	863 / 672	150 / 117	39 / 34	7 / 6
`Applicant`	932 / 677	162 / 122	44 / 38	8 / 7
`Admitted`	769 / 605	135 / 108	38 / 31	6 / 5
`Undergraduate`	1091 / 879	190 / 155	43 / 36	7 / 6
`Graduate`	1209 / 989	206 / 162	51 / 46	8 / 7
Overall	**1014 / 786**	**177 / 139**	**43 / 37**	**7 / 6**

Figure 3. Word clouds for the posts (in PT-BR). (left) Distribution of terms found in post titles. (right) Distribution of terms found in post bodies. The bigger the word in the figure, the higher its frequency in the posts' titles or bodies.

Table 6. Comments values and length by subreddit.

subreddit	Comments	Unique Users	Body (Chars) (Mean / Median)	Body (Words) (Mean / Median)
r/askacademico	1,362	528	348 / 198	60 / 35
r/faculdadeBR	7,397	2,784	298 / 158	52 / 28
r/USP	6,921	2,273	29 / 29	5 / 6
Overall	**15,680**	**5,229**	**306 / 164**	**53 / 29**

related words like college (*faculdade*), course (*curso*), USP, and master's (*mestrado*).

5.2. Comments

The dataset comprises **15,680** comments associated with the collected posts, involving 5,229 unique users (Tab. 6). The number of unique users is significantly smaller than the number of comments, suggesting that individual users tend to comment multiple times. Similarly to the posts distribution, comments are concentrated in the *r/faculdadeBR* subreddit and are predominantly associated with the *undergraduate* level. Among these, 370 comments were already marked as deleted in the original collection, with their content replaced with "`[deleted]`" by the REDDIT API.

Regarding textual content, comments are considerably shorter than posts, with an overall mean of 306 characters (53 words) and a median of 164 characters (29 words), as shown in Tab. 6. As with the posts, slight variations are observed across subreddits: *r/askacademico* present longer comments, while *r/USP* has notably shorter texts. The consistent gap between mean and median suggests the presence of longer comments that increase the average length, as well as the influence of deleted content.

6. Applications

The `BrStudentMH` dataset can provide a foundation for developing automated systems designed to safeguard the digital environment of university students. One primary application is the training of supervised machine learning models to accurately identify and filter potentially harmful content — such as posts or comments that might exacerbate existing mental health conditions or trigger emotional distress. The necessity for such tools is supported by recent research indicating a high demand among students for online mental health information and a tendency to utilize digital psychological support [Vornholt and De Choudhury 2021].

Beyond individual and institutional monitoring, the dataset also enables the study of informal digital peer support networks. By linking posts to their respective comments, `BrStudentMH` allows researchers to analyze not only how support is sought, but also how it is offered and reciprocated within online care networks. This application is particularly relevant as research indicates that young people derive significant mental health benefits from the immediacy of friendship and peer-based solidarity found in online spaces [Byron and McDaid 2026]. Utilizing this dataset, social scientists and psychologists can investigate where informal support originates and how these spontaneous care networks help Brazilian students navigate the specific pressures of the university environment, potentially serving as a model for enhancing existing institutional support frameworks.

7. LLM Performance

For the two performed categorization tasks — whether the post is related to mental health and the academic level reflected in the text —, we analyzed the models' performance from two perspectives. The first is the inter-model agreement rate, that is, how often two models produce the same result (Tables 7 and 8). The second examines the agreement between a model's output and the majority vote, including cases of disagreement (Tab. 9). Majority vote disagreement is defined as a situation in which all other three models converge on the same category, while the model under evaluation selects a different one.

Regarding the first perspective, Tab. 7 shows that DeepSeek V3 and Grok 3 achieved the highest agreement rate for the mental health categorization (94.65%). In contrast, the lowest agreement rate was observed between DeepSeek V3 and Gemini 2.5 Pro (84.84%). In the case of the academic level categorization (Tab. 8), the highest agreement rate was observed with Grok 3 and Gemini 2.5 Pro (86.53%), while the lowest was observed with DeepSeek V3 and ChatGPT 4o (79.22%).

With respect to the (dis)agreement rate between a model's output and the three other models' decision (second perspective), one can observe from Tab. 9 (upper portion) that Grok 3 and Gemini 2.5 Pro are, respectively, the models with the highest and lowest agreement with the majority vote in the case of mental health categorization (97.27% and 89.71%). Consistently, these models also exhibit the fewest and most disagreements with the majority vote, respectively.

In the case of academic level categorization (Tab. 9 (bottom portion)), an inversion in Gemini 2.5 Pro's performance is observed: the model's outputs are now more aligned with the majority vote (92.36%), and it rarely disagrees with the other three models (2.65%). Regarding the Grok 3's performance, although it shows the highest alignment with Gemini 2.5 Pro (Tab. 8), it ranks third in terms of agreement with the majority vote (89.08%). Finally, ChatGPT 4o presents the lowest agreement rate with the majority vote among all models (only 86.21%).

Table 7. Inter-model agreement rate for Mental Health categorization

Mental Health	DeepSeek V3	ChatGPT 4o	Grok 3	Gemini 2.5 Pro
DeepSeek V3	100.00%			
ChatGPT 4o	92.11%	100.00%		
Grok 3	**94.65%**	93.60%	100.00%	
Gemini 2.5 Pro	**84.84%**	86.94%	88.08%	100.00%

Table 8. Inter-model agreement rate for Academic Level categorization

Academic Level	DeepSeek V3	ChatGPT 4o	Grok 3	Gemini 2.5 Pro
DeepSeek V3	100.00%			
ChatGPT 4o	**79.22%**	100.00%		
Grok 3	80.81%	80.70%	100.00%	
Gemini 2.5 Pro	84.09%	81.97%	**86.53%**	100.00%

Table 9. Agreements and disagreements with the majority vote (MV)

Category	Metric	DeepSeek V3	ChatGPT 4o	Grok 3	Gemini 2.5 Pro
Mental Health	MV Agreement	94.20%	96.13%	**97.27%**	**89.71%**
	3:1 MV Disagreement	3.07%	2.54%	**0.70%**	**8.94%**
Academic Level	MV Agreement	89.50%	**86.21%**	89.08%	**92.36%**
	3:1 MV Disagreement	6.36%	**7.53%**	4.77%	**2.65%**

8. Challenges and Limitations

Data collection is a critical aspect of constructing a dataset. With REDDIT, there is an additional challenge: subreddits are defined by their titles, which may not accurately reflect the content they predominantly share. Therefore, identifying relevant subreddits depends on searching for keywords in posts and verifying which subreddit each post belongs to. In this regard, searches for mental health-related terms may return posts that do not necessarily align with the topic, often due to colloquial usage or differing semantics. Also, note that our dataset focus on Brazilian university students who use Reddit, thus the findings may be influenced by characteristics and biases specific to this population.

Although we have used different LLMs, it is important to note that the categorization of academic levels was performed automatically and based exclusively on the textual content of the posts, which may not accurately reflect the actual level in all cases. The pseudonymization process, in turn, combined both automatic and manual inspections. However, the informal writing style commonly found on social platforms (e.g., abbreviations) poses challenges that may not have been fully addressed in this work.

9. Conclusion

This paper introduces `BrStudentMH`, a dataset of 943 posts and 15,680 comments in PT-BR focused on Brazilian higher education students' mental health, constructed through data extraction from REDDIT, LLM-based categorization, and pseudonymization. `BrStudentMH` was developed to assist in a diversity of studies and applications. For example, it enables the development of automated systems for detecting harmful content and supports research into online peer support dynamics. It also enables analysis of how students seek and provide support in digital environments, offering valuable insights to improve institutional mental health strategies.

Future work includes generating keywords based on the central themes discussed in both posts and comments, going beyond the original text to produce more representative descriptors of the content. Additionally, applying data mining techniques to extract and include attributes not present in all posts, but informative when available, could further enrich the dataset. Examples include age, university, city, organization, medications, and other context-specific information related to the university environment. Another promising direction is the investigation of NER techniques tailored to online environments, particularly for handling slang, abbreviations, and similar linguistic variations.

10. Acknowledgments

This research was supported by the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant 144405/2025-3), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Finance Code 001), São Paulo Research Foundation (FAPESP, grant No. 2024/13328-9), and University of São Paulo (PUB-USP and PRPI Ordinance No. 1032, Process 22.1.09345.01.2).

Artificial Intelligence Usage

ChatGPT and Grammarly were used for grammatical revision of this manuscript text. The authors assume full responsibility for the content of the work.

References

An, J., Kim, J., Sunwoo, L., Baek, H., Yoo, S., and Lee, S. (2025). De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained bert. *BMC Medical Informatics and Decision Making*, 25(1):82.

Byron, P. and McDaid, L. (2026). Informal digital peer support for mental health: understanding the digital support practices of LGBTQ+ young people in Australia. *Culture, Health & Sexuality*, 28(3):321–338.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., and Goharian, N. (2018). SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 1–10.

Croes, E. A., Antheunis, M. L., Van Der Lee, C., and De Wit, J. M. (2024). Digital confessions: The willingness to disclose intimate information to a chatbot and its impact on emotional well-being. *Interacting with Computers*, 36(5):279–292.

da Silva Nascimento, R., Parreira, P., dos Santos, G., and Guedes, G. P. (2018). Identificando sinais de comportamento depressivo em redes sociais. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, pages 235–240, Porto Alegre, RS, Brasil. SBC.

Guillou, P. (2021). ner-bert-large-cased-pt-lenerbr: Bert large ner model for portuguese legal domain (lener-br). `https://huggingface.co/pierreguillou/ner-bert-large-cased-pt-lenerbr`. Hugging Face model. Accessed: 2026-04-27.

Guthold, R., Carvajal-Velez, L., Adebayo, E., Azzopardi, P., Baltag, V., Dastgiri, S., Dua, T., Fagan, L., Ferguson, B. J., Inchley, J. C., et al. (2023). The importance of mental health measurement to improve global adolescent health. *Journal of Adolescent Health*, 72(1):S3–S6.

Herculano, A., de Paula, T.-H., Fernandes, D., and Rego, A. (2024). DepreRedditBR: Um conjunto de dados textuais com postagens depressivas no idioma português brasileiro. In *Anais do VI Dataset Showcase Workshop*, pages 77–90, Porto Alegre, RS, Brasil. SBC.

Klann, V. and Sousa, E. (2024). Brsuicides: Conjunto de dados para análise de casos diários de suicídio no brasil. In *Anais do VI Dataset Showcase Workshop*, pages 12–22, Porto Alegre, RS, Brasil. SBC.

Li, Q., Li, J., and Fan, Y. (2025). Addressing mental health in university students: a call for action. *Frontiers in Public Health*, 13:1614999.

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., and Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *J Med Internet Res*, 22(10):e22635.

Ruihua, L., Hassan, N. C., Qiuxia, Z., Sha, O., and Jingyi, D. (2025). A systematic review on the impact of social support on college students' wellbeing and mental health. *Plos one*, 20(7):e0325212.

Sampath, K. and Durairaj, T. (2022). Data set creation and empirical analysis for detecting signs of depression from social media postings. In Kalinathan, L., R., P., Kanmani, M., and S., M., editors, *Computational Intelligence in Data Science*, pages 136–151, Cham. Springer International Publishing.

Santos, W. R. d., de Oliveira, R. L., and Paraboni, I. (2023). Setembrobr: a social media corpus for depression and anxiety disorder prediction. *Lang. Resour. Eval.*, 58(1):273–300.

Times Higher Education (2025). Latin america university rankings 2026. `https://www.timeshighereducation.com/world-university-rankings/2026/latin-america-university-rankings`. Accessed: 11 dez. 2025.

Turcan, E. and McKeown, K. (2019). Dreaddit: A Reddit dataset for stress analysis in social media. In Holderness, E., Jimeno Yepes, A., Lavelli, A., Minard, A.-L., Pustejovsky, J., and Rinaldi, F., editors, *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Vornholt, P. and De Choudhury, M. (2021). Understanding the role of social media–based mental health support among college students: Survey and semistructured interviews. *JMIR Ment Health*, 8(7):e24512.

Voss, C., Shorter, P., Weatrowski, G., Mueller-Coyne, J., and Turner, K. (2023). A comparison of anxiety levels before and during the covid-19 pandemic. *Psychological Reports*, 126(6):2669–2689. PMID: 35503814.

Xu, Y., Fang, Z., Lin, W., Jiang, Y., Jin, W., Balaji, P., Wang, J., and Xia, T. (2025). Evaluation of large language models on mental health: from knowledge test to illness diagnosis. *Frontiers in Psychiatry*, 16:1646974.