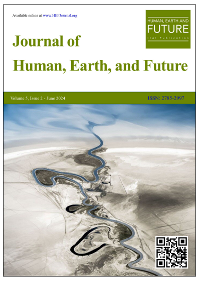
Available online at www.HEFJournal.org

Journal of
Human, Earth, and Future

ISSN: 2785-2997

Vol. 5, No. 2, June, 2024



Exploring Mental Stress Expressions in Online Communities: A Subreddit Analysis

Tran Anh Tuan [1, 2*], Nguyen Huu Nghia [3], Tran Dai An [4], Dao Thi Thanh Loan [5]

[1] *School of Informatics, Walailak University, Nakhon Si Thammarat, 80160, Thailand.*

[2] *Informatics Innovation Center of Excellence (IICE), Walailak University, Nakhon Si Thammarat, 80160, Thailand.*

[3] *Quang Binh Medical College, Quang Binh, 47100, Vietnam.*

[4] *Dong Thap Provincial Center for Disease Control, Dong Thap, 81000, Vietnam.*

[5] *Faculty of Foundational Sciences, Dak Lak College of Pedagogy, Dak Lak, 63000, Vietnam.*

Received 09 February 2024; Revised 16 May 2024; Accepted 22 May 2024; Published 01 June 2024

Abstract

Objectives: This study aims to comprehensively explore trends, sentiments, and visualization of mental stress expressions in online communities, focusing on discussions within subreddits on the social media platform Reddit. *Methods/Analysis*: Advanced text analysis and statistical techniques are employed to achieve the study's objectives. The research utilizes natural language processing (NLP) methods, sentiment analysis, and topic modeling to unravel the intricate layers of mental stress expressions found in posts across diverse subreddits. Additionally, engagement metrics, such as Redditors' scores and the number of comments, are analyzed to discern distinctive information and patterns of interest. *Findings*: The research sheds light on prevalent trends, sentiments, and themes related to mental stress in online conversations before and after January 2020. The findings provide valuable insights into patterns of exciting topics, shared experiences of stress, coping mechanisms, and the significant role of virtual communities in offering support and understanding. *Novelty/Improvement*: The novelty lies in applying advanced text analysis techniques, including sentiment analysis with the majority voting method combining different machine learning techniques and topic modeling with semantic networks, to gain a deeper understanding of the dynamics of mental stress expressions in online communities. The research explores current patterns and distinguishes itself by examining temporal variations in stress-related posts and their correlation with engagement metrics, offering an innovative perspective on mental health discussions in the digital age.

Keywords: Mental Stress; Sentiment Analysis; Public Opinion; Subreddits; Natural Language Processing.

1. Introduction

Mental stress is a complex and widespread aspect of the human experience that has gained increasing attention in the digital age. It encompasses a range of emotions, from mild apprehension to severe anxiety, often stemming from the difficulties of human relationships, financial pressures, and existential uncertainties [1]. Social media platforms, including Reddit, provide important arenas where people from different backgrounds can come together to express their concerns, seek comfort, and share coping strategies [2]. Subreddits are dedicated to specific themes or subjects on the social media platform Reddit, where users can openly express their daily experiences and emotions. On social media, stressed users usually post about exhaustion, losing control, increased self-focus, and physical pain, while non-stressed users focus on topics such as breakfast, family time, and travel [3]. Moreover, features of social media (e.g., comments,

* Corresponding author: tran.tu@wu.ac.th

http://dx.doi.org/10. 28991/HEF-2024-05-02-01

likes) highly reflect the users' mental stress [4, 5]. Hence, analyzing public opinion and sentiment within these online communities can offer essential insights into the prevalence and dynamics of mental stress, inform targeted mental health interventions and support strategies, and provide valuable insights into how mental stress is perceived and managed within distinct online communities. On the other hand, the COVID-19 pandemic has immensely impacted mental health and well-being, exacerbating existing stressed users and introducing new challenges [6], consequently prompting people to turn to online platforms for support, information, and connection [7]. Therefore, it is essential to investigate mental stress within the context of the pandemic to provide a window into the evolving landscape of human emotions and behaviors during times of crisis.

Recent Natural Language Processing (NLP) techniques support analyzing textual information (e.g., sentiment analysis, topic modeling) that could add value to the posts/comments and provide valuable insights for users. Studies demonstrated machine learning methods' effectiveness in accurately detecting stress and non-stress posts within subreddit communities. For instance, Saha et al. [8] proposed machine learning classifiers and a lexicon of linguistic markers to identify and contextualize minority stressors, which can improve knowledge relating to the mental health disparities of LGBTQ+ populations from subreddit 'lgbt'. In Febriansyah et al.'s research [9], the authors aimed to detect whether social media users were under stress from a Reddit dataset (5 categories) using machine learning classifiers. Shen & Rudzicz [10] focused on detecting anxiety disorders through personal narratives collected from Reddit. The authors built a dataset of typical and anxiety-related posts and applied various techniques to classify posts related to binary levels of anxiety accurately. Additionally, some research has focused on identifying the specific topics that users discuss within these communities or investigating the factors that influence subreddit posts [10, 11]. Naseem et al. [12] identified discussion communities on Reddit that frequently mentioned COVID-19-related terms and applied the Non-negative Matrix Factorization topic modeling algorithm to extract topics discussed by different communities. Gao et al. [13] applied topic modeling techniques (Latent Dirichlet Analysis, Neural Topic Model with Knowledge Distillation, and Embedded Topic Model) to analyze maternal health topics, concerns, and questions expressed in online communities on social networking sites, specifically Reddit. Stevens et al. [14] investigated conversation topics in the most popular subreddit for LGBTQ+ youth in the COVID-19 pandemic effects on mental health, including increased anxiety and depressive symptoms among adults, by using natural language processing methodologies. Zhu et al. [15] used natural language processing techniques and statistical methods to thematically analyze mental health support groups on Reddit during the COVID-19 pandemic. Papakyriakopoulos [16] investigated the relationship between social media reaction mechanisms (upvotes, downvotes) and political rhetoric in user discussions on Reddit. These studies collectively highlight the importance of leveraging analytical techniques to gain a deeper understanding of the dynamics of mental stress and online interactions within digital platforms. There is no study explicitly focusing on the comprehensive analysis of trends, sentiments, or visualizations of mental stress.

This study aims to explore public opinion about mental stress through text analysis of posts within twelve selected subreddits before and after January 2020, seeking to uncover the emotional expressions, attitudes, and subjectivity associated with mental stress discussions by advanced text analysis techniques (e.g., sentiment analysis with the majority voting method combining different machine learning techniques and topic modeling with semantic networks) and statistics techniques (qualitative and quantitative methods). Moreover, we investigate any temporal patterns in stress-related posts to understand how stress experiences may vary over time within these online communities and any factors of engagement metrics that affect how much stress is across subreddits via polarity scores. To gain a comprehensive understanding of public perceptions and emotional experiences surrounding mental stress in subreddit communities, we state five research questions (RQs):

- ***RQ1.*** Do the engagement metrics (e.g., Redditors' scores, number of comments) of the posts from the crowd offer distinctive information to the public?
  - *RQ1.1.* Can we find a strong interest in discussions from the engagement metrics?
- ***RQ2.*** Does sentiment analysis of the posts provide any insight into stress levels for the public?
- ***RQ3.*** Can we identify trends and fluctuations in stress levels from posts?
- ***RQ4.*** Is there any correlation between polarity scores and the engagement metrics (Redditors' scores, number of comments) in each stress level or not?
- ***RQ5.*** Can we leverage NLP-driven techniques to find important themes?
  - *RQ5.1.* What is the theme based on the highly frequent words for each subreddit?
  - *RQ5.2.* What is the theme based on semantic networks for each stress level?

The findings from this study can enhance comprehension of Redditors' stress across subreddits, providing valuable implications for researchers developing health interventions. Additionally, the research diagram of this study can be applied to other health conditions.

2. Proposed Methodology

Our analysis comprises several distinct sequential stages, including (1) *Data collection*, (2) *Data pre-processing*, and (3) *Data modelling*. Figure 1 provides a visual representation of the complete methodology process. In the subsequent sections, we will elaborate on each stage, providing a detailed explanation.

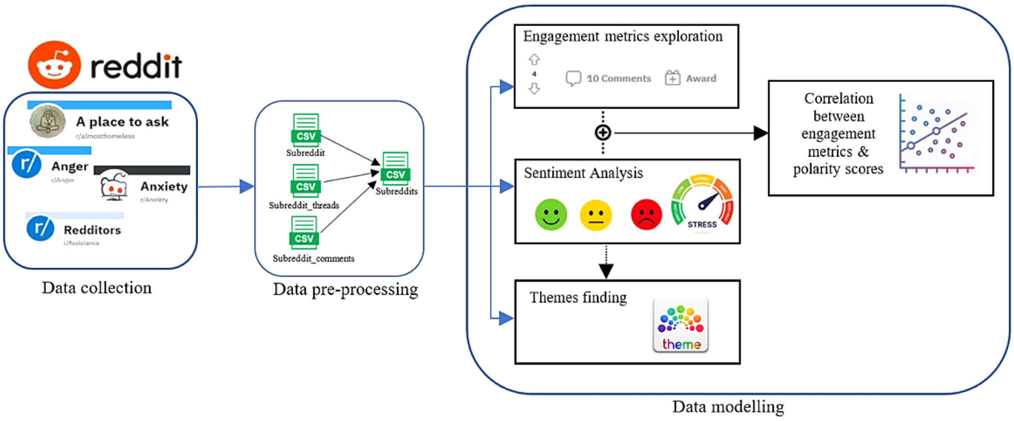


Figure 1. Methodology diagram

2.1. Data Collection

Posts were collected from subreddits, namely *r/almosthomeless*, *r/anger*, *r/anxiety*, *r/assistance*, *r/cptsd*, *r/depression*, *r/food_pantry*, *r/homeless*, *r/mentalhealth*, *r/ptsd*, *r/relationships*, and *r/stress*. The *RedditExtractoR* package was used in this study [17].

2.2. Data Pre-processing

The collected posts from each subreddit were saved in three separate CSV files: *Subreddit.csv*, *Subreddit_threads.csv*, and *Subreddit_comments.csv*. These files were combined based on their timestamps and retained under the corresponding subreddit names. All the files from the various subreddits were concatenated into one unified dataset. The structure of the final dataset consisted of columns: *time stamp* (timestamp), *text*, *Redditors' scores* (Redditor_scores), *number of comments* (No_comments), *total received awards* (Total_received_awards), *golds*, *up ratio*, and *subreddit*. In the next step of this phase, stopwords, punctuation, and URLs were removed from the text using a function in the quanteda package in R [18].

2.3. Data Modelling

This subsection is divided into five separate parts, each explaining the analysis to address one research question (RQ):

Engagement metrics exploration for the posts:

We employ two steps to examine the distribution of the ratings in the data collection. In the first step, we aggregate the total posts in each subreddit along with Redditors' scores, number of comments, total received awards, golds, and up ratio for each post. These are segregated based on timestamps before and after January 2020. In the second step, we delve into the correlations between the Redditors' scores and the number of comments within each subreddit. This exploration aims to uncover engagement patterns, with a particular focus on identifying high positive correlations, which would indicate a strong interest in discussions within the subreddit. We represent the patterns on scatter graphs with Spearman's rank correlation coefficient by using the ggscatter function in R [19].

Sentiment analysis:

Sentiment analysis, also known as opinion mining, is a technique used to determine opinions, emotions, subjectivity, and attitudes expressed in natural language text [20–22]. In the context of stress analysis, sentiment analysis is used to understand the emotions and sentiments in stress-related discussions. The knowledge can improve mental health support, emotional well-being, and targeted stress management interventions within online communities. The current work uses a majority voting method, in which three various libraries are employed, namely TextBlob [23], VADER [24], and Flair [25], to determine the stress level of posts within subreddits (TextBlob, VADER, and Flair return polarity and polarity score). Figure 2 depicts the model for determining stress levels.

In the ensemble determiner for stress levels in Figure 2, the majority voting method compares the returned results from the models, selecting the polarity that occurs most frequently among the models. Then, the polarity scores of the most frequent models are used to calculate an average. In cases where the returned results of the three models are different, the polarity scores of the three models are used to calculate the summation score for the ensemble determiner.

A post with a negative score is considered a stress post. The average polarity score is mapped to the stress level scales [26], including *very low stress* (0-20), *low stress* (21-40), *moderate stress* (41-60), *high stress* (61-80), and *very high stress* (81-100). It is worth noting that the average polarity score has an interval of [-1.0, 1.0]. We use Google Colab to deploy the three models to calculate polarity scores, after which the results are kept in the final dataset in CSV format for the other phases.

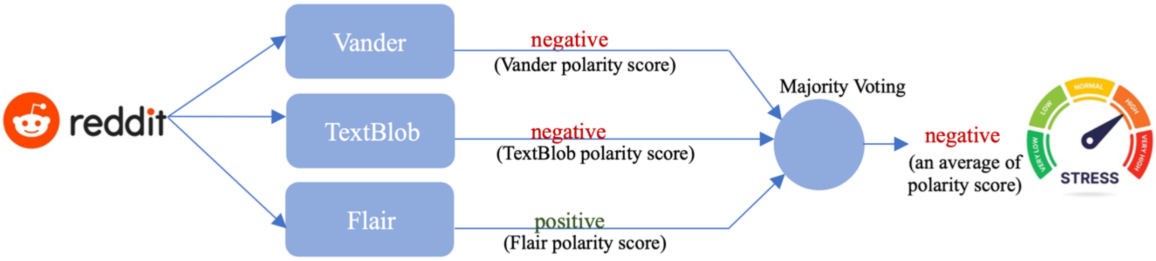


Figure 2. An ensemble determiner for stress level

Temporal analysis for stress-related posts:

Temporal analysis for stress levels involves examining the fluctuations and patterns of stress levels within the subreddit community over time. We can gain valuable insights into how stress experiences vary throughout the temporal domain by examining stress data across different time intervals (years). We use the stress scores for each post returned in the sentiment analysis phase and the respective timestamps in the dataset to conduct temporal analysis. Then, the results are used to visualize through timestamps to explore the pattern of stress levels in specific years. These findings will be crucial in shaping targeted support strategies, interventions, or stress management programs, aligning them with the specific stress challenges the community faces at different times. The graph was represented using the alluvial package in R [27].

Correlation between polarity scores and the engagement metrics of stress-related posts:

In this section, we employ two distinct approaches to explore the correlation between polarity scores and the engagement metrics of stress-related posts. The first approach investigates correlations between polarity scores and Redditors' scores, as well as correlations between polarity scores and the number of comments within each stress level. This analysis provides insights into how stress levels vary at different times and how user engagement relates to stress-related discussions. We utilize Spearman's rank correlation coefficient for these correlations and represent them using scatter graphs created with the ggscatter function in R [19]. The second approach seeks to illustrate which factors among the engagement metrics influence polarity scores (square root transformed) across all posts in the stress group. This analysis is conducted using linear regression, with the following variables investigated: (1) *Redditors' scores*, (2) *number of comments*, (3) *total received awards*, (4) *golds*, and (5) *Up ratio*. We build a multivariable model using a stepwise approach to select the final model. Univariable models are screened, and those with a significance level of p<0.20 are retained as candidates for the multivariable models. All statistical analyses are conducted using R [28].

Finding themes for stress-related posts:

Based on subreddits and five distinct stress levels, we employ two methods utilizing natural language processing (NLP) to identify the primary themes within stress-related posts. In the first method, aimed at discovering themes within subreddits, we calculate scores for words (bi-grams) within each subreddit. These scores reflect differential occurrences across different categories, specifically posts before and after January 2020. The scoring system comprises four values: chi-square (χ^2), p-value, and the frequencies before and after January 2020. We concatenate the top five words based on their χ^2 scores in each subreddit. The resulting list of chosen words is then visually represented on a graph and is used to synthesize a theme for each subreddit. In the second method, focused on identifying themes within stress levels, we examine the co-occurrence frequencies of words within each stress level. Words with high frequencies are selected, and a theme for each stress level is derived. We use functions within the quanteda package in R [18] to calculate scores, create graphs, and synthesize themes.

3. Results

3.1. RQ1. Do the engagement metrics (e.g., Redditors' scores, number of comments) of the posts from the crowd offer distinctive information to the public?

Of the 136,638 posts across twelve subreddits within more than twelve years (May 2011 to July 2023), 64,170 (46.96%) and 72,468 (53.04%) posts and comments were expressed by Redditors before and after January 2020, respectively. Table 1 describes characteristics distribution of posts in each subreddit split by date.

The subreddit with the highest number of posts (38.74%) was *relationships*. Within the relationships subreddit, there were 44,078 posts before January 2020 and 8,853 posts after. Subreddits that accounted for more than 10% of the total of posts and comments included *CPTSD* at 17.68% (1,700 posts before January 2020 and 22,464 posts after), *Anxiety* at 11.68% (5,785 and 10,178), and *Assistance* at 10.09% (3,019; 10,770). All other subreddits had percentages below 10%. Redditors' scores for a post ranged from 0 to 16,029 (median [IQR]: 3[2-13]). Four intervals ([0, 1], [2, 3], [4, 21], and [22, 16029]) of the Redditors' scores accounted for the following percentages of total posts: 24.74% (15,043 posts before January 2020 and 18,761 posts after), 27.41% (16,106; 21,346), 29.11% (18,404; 21,373), and 18.74% (14,617; 10,988).

The number of comments for a post ranged from 0 to 10,137 (median [IQR]: 0[0-1]). The majority (63.36%) of total posts had no comments (42,410 posts before January 2020 and 44,163 posts after). A total of 36.64% of posts (21,760 and 28,305 posts before and after January 2020, respectively) had some comments ranging from 1 to 10,137 (step 1). The total received awards for a post ranged from 0 to 41 (median [IQR]: 0[0-0]). The majority (87.19%) of total posts had no awards (62,832 posts before and 56,304 posts after January 2020). In 12.81% of posts (1,338 and 16,164 posts before and after January 2020, respectively), the number of comments ranging from 1 to 41. Golds for a post ranged from 0 to 3 (median [IQR]: 0[0-0]). The majority (99.19%) of total posts had no golds (63,936 posts before January 2020 and 71,592 posts after). In 0.81% of posts (234 and 876 posts before and after January 2020, respectively). the number of comments ranging from 0.5 to 3.0. Up ratio for a post ranged from 0 to 1 (median [IQR]: 0[0-0]). The majority (96.59%) of total posts had no up-ratio (62,693 posts before January 2020 and 69,288 posts after). In a total of 3.41% of posts (1,477 and 3,180 posts before January 2020 and later, respectively), the number of comments ranged from 0.7 to 1.0.

Table 1. Descriptive characteristics of posts in the subreddit. The breakdown of the total sample size into the date of a timestamp is based on the date (January 2020) of the timestamp

Characteristics	All posts (n=136,638)	Date < January 2020 No. of posts (n=64,170)	Date ≥ January 2020 No. of posts (n=72,468)
Subreddit (%)			
almosthomeless	3,740 (2.74%)	982	2,758
Anger	2,565 (1.88%)	705	1,860
Anxiety	15,963 (11.68%)	5,785	10,178
Assistance	13,789 (10.09%)	3,019	10,770
CPTSD	24,164 (17.68%)	1,700	22,464
depression	10,878 (7.96%)	4,550	6,328
Food_Pantry	2,021 (1.48%)	1,474	547
homeless	5,894 (4.31%)	584	5,310
mental health	217 (0.16%)	194	23
ptsd	3,288 (2.41%)	931	2,357
relationships	52,931 (38.74%)	44,078	8,853
Stress	1,188 (0.87%)	168	1,020
Redditors' score(s) (%)			
0-1	33,804 (24.74%)	15,043	18,761
2-3	37,452 (27.41%)	16,106	21,346
4-21	39,777 (29.11%)	18,404	21,373
22 - 16,029	25,605 (18.74%)	14,617	10,988
No. of comments (%)			
0	86,573 (63.36%)	42,410	44,163
≥1 (1-10,137)	50,065 (36.64%)	21,760	28,305
Total received awards (%)			
0	119,136 (87.19%)	62,832	56,304
≥1 (1-41)	17,502 (12.81%)	1,338	16,164
Golds (%)			
0	135,528 (99.19%)	63,936	71,592
≥ 0.5 (0.5-3.0)	1,110 (0.81%)	234	876
Up-ratio (%)			
0	131,981 (96.59%)	62,693	69,288
≥ 0.7 (0.7-1.0)	4,657 (3.41%)	1,477	3,180

RQ1.1. Can we find a strong interest in discussions from the engagement metrics?

All twelve subreddits exhibited vibrant discussions across their respective total posts, covering a comprehensive range of dates in the dataset. These engaging discussions were indicated by positive correlations between Redditors' scores and the number of comments, with statistically significant results (all $p < 0.05$). The most substantial correlation coefficient was observed in the *Anger* subreddit, registering at R = 0.76 with $p < 2.2e\text{-}16$.

When analyzing the distribution of posts based on the date, explicitly distinguishing between periods before and after January 2020, several subreddits displayed a remarkable surge in discussions. Notably, the *mental health* and *stress* subreddits experienced considerable spikes. For the *mental health* subreddit, the correlation coefficient increased substantially from 0.21 to 0.75 ($p < 0.001$) after January 2020, while the *stress* subreddit saw its coefficient climb from 0.058 to 0.57 ($p < 2.2e\text{-}16$). The *Anger* subreddit also experienced a moderate increase in discussions, with its correlation coefficient rising from 0.68 to 0.78 (both $p < 2.2e\text{-}16$). Conversely, the *ptsd* subreddit witnessed a decline in discussions with (both $p > 0.05$). The remaining subreddits maintained consistent correlation coefficients (Figure A1; Appendix I).

3.2. RQ2. Does sentiment analysis of the posts provide any insights into stress levels for the public?

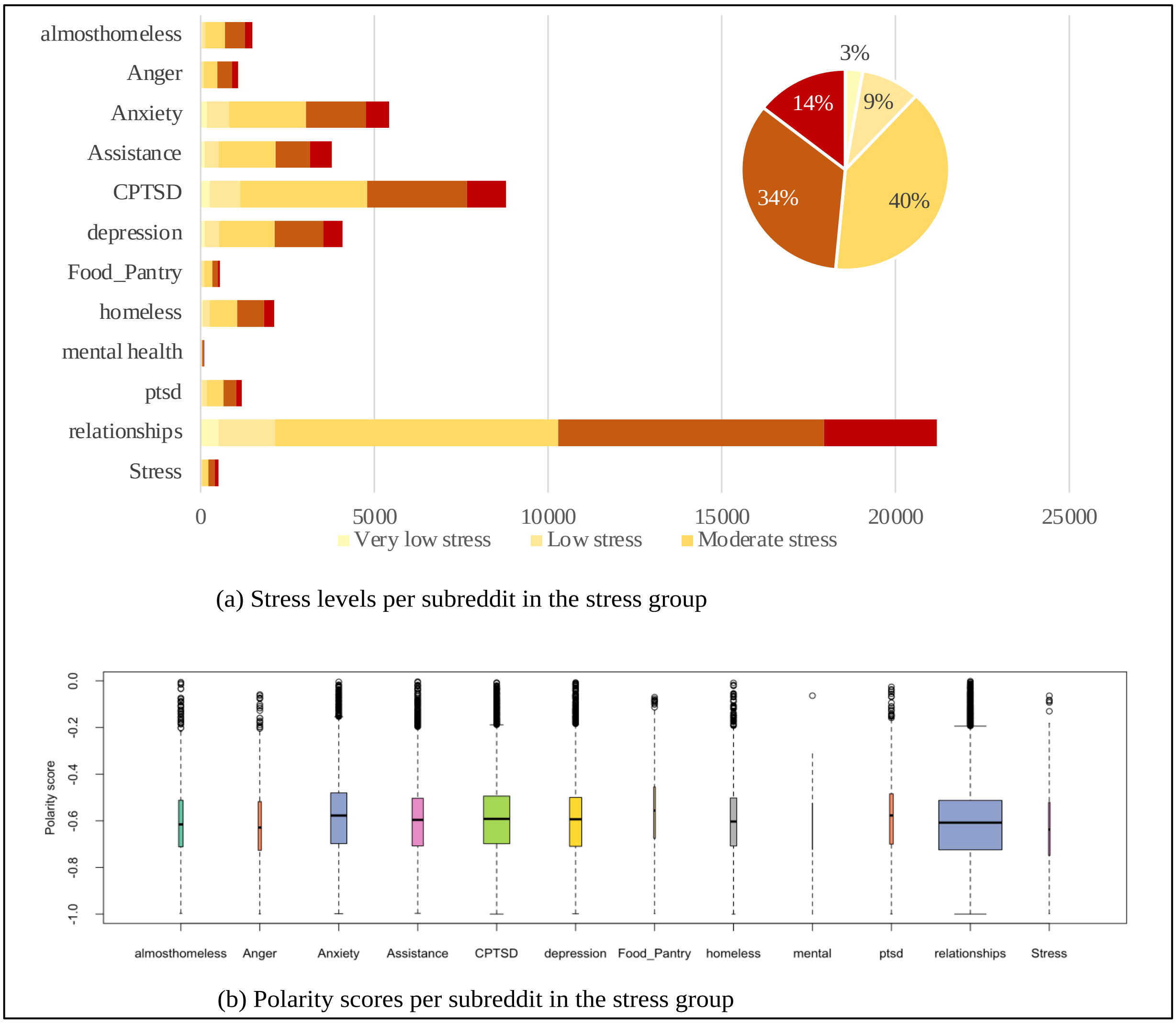
As mentioned in Section 2.3.2, we deployed an ensemble determiner with three libraries to assess stress levels for posts in subreddits. The ensemble determiner provided a polarity score, which was then mapped to non-stress and stress levels. After deploying the ensemble determiner to all posts in the dataset, the results for each subreddit were categorized into non-stress and five stress levels, as depicted in Figure 3 and Table 2.

The number of posts in the data was classified into two groups: non-stress and stress, accounting for 86,350 (63.20%) and 50,288 (36.80%) of the total posts, respectively. In the non-stress group, the *relationships* subreddit had the highest number of posts, comprising 36.77% of the non-stress group. The second highest number of posts in this group was in the *CPTSD* subreddit, accounting for 17.81%. The remaining subreddits had fewer than 11,000 posts each. All posts in the non-stress group had mean polarity scores ranging from 0.461 to 0.535, with a slight standard deviation (SD < 0.31). In the stress group, the *relationships* and *CPTSD* subreddits also had the highest number of posts, constituting 42.12% and 17.47%, respectively. The mean polarity score for posts in this group was < -0.5, with a slight standard deviation (SD < 0.2) (Table 2). There were five stress levels, and the *high-stress* and *moderate-stress* levels were the most prevalent across all twelve subreddits, accounting for 34.00% and 40.00% of the stress group, respectively (Figure 3 and the last five columns of Table 2). Furthermore, the polarity scores for the stress group ranged from -1 to -0.001 across the twelve subreddits (Figure 3). This information provides valuable insights into the stress levels and sentiment patterns within the examined subreddits.

Table 2. Descriptive statistic for stress and non-stress posts in each subreddit

Characteristics	Non-stress Group			Stress Group							
	No. posts (n=86,350)	Polarity score		No. posts (n=50,288)	Polarity score		No. of posts in stress levels				
		Mean	SD		Mean	SD	VL	L	M	H	VH
almosthomeless	2,259	0.461	0.246	1,481	-0.616	0.177	32	103	563	565	218
Anger	1,484	0.489	0.230	1,081	-0.628	0.177	18	77	387	419	180
Anxiety	10,535	0.535	0.244	5,428	-0.581	0.188	183	631	2,216	1,727	671
Assistance	10,011	0.465	0.305	3,778	-0.596	0.188	116	400	1,647	989	626
CPTSD	15,379	0.520	0.239	8,785	-0.591	0.184	263	880	3,659	2,859	1,124
depression	6,792	0.491	0.250	4,086	-0.598	0.186	117	409	1,604	1,398	558
Food_Pantry	1,470	0.528	0.228	551	-0.5594	0.190	22	81	239	151	58
homeless	3,773	0.499	0.251	2,121	-0.603	0.183	60	201	784	776	300
mental health	118	0.470	0.251	99	-0.632	0.162	1	4	40	37	17
ptsd	2,104	0.504	0.242	1,184	-0.585	0.184	31	140	483	373	157
relationships	31,747	0.470	0.251	21,184	-0.613	0.180	515	1,633	8,138	7,657	3,241
Stress	678	0.465	0.235	510	-0.637	0.179	8	35	176	195	96

Note: VL, L, M, H, and VH are very low stress, low stress, moderate stress, high stress, and very high stress, respectively.



(a) Stress levels per subreddit in the stress group

(b) Polarity scores per subreddit in the stress group

Figure 3. Stress levels and polarity scores in each subreddit of the stress group

3.3. RQ3. Can we identify trends and fluctuations in stress levels from posts?

Among the 50,288 stress-related posts depicted in Figure 4, the width of the ribbons visually represents the prevalence of posts across different subreddit categories. Notably, the *relationships* and *CPTSD* subreddits stand out prominently. Similarly, when examining stress levels, moderate- and high-stress levels emerge as the most prevalent, highlighted by the varying widths of the ribbons.

Before January 2020 (from 2011 to 2019), the number of stress-related posts exhibited variations within different subreddit categories. In the earlier years of this period (2011 - 2019), the majority of stress-related posts from Redditors were found in the *relationships* subreddit. However, the number of posts in this subreddit experienced a slight decline from 2017 onwards. Concurrently, the number of posts in three other subreddits (*CPTSD, Anxiety*, and *depression*) has increased since 2020, predominantly falling within the moderate, high, and very high-stress levels (as illustrated in Figure 4).

The count of subreddits featuring stress-related posts escalated from one in 2011 to 12 by the end of 2019. During this period (2011 - 2019), the number of posts across the five stress levels increased annually. Particularly noteworthy is the peak number of posts reached in each stress level in 2015 (171 posts for very low stress level, 581 posts for low stress, 2,743 posts for moderate stress, 2,698 posts for high stress, and 1,089 posts for very high stress level). Over the subsequent four years, from 2020 to 2023 (after January 2020), subreddits hovered around 11 categories. The zenith of post activity during this timeframe was achieved in 2021 (92 posts for very low stress level, 275 posts for low stress, 1,162 posts for moderate stress, 853 posts for high stress, and 350 posts for very high stress level), as presented in Table 3. The flow of ribbons from one subreddit category to another demonstrates how the stress levels change or shift over time within and between different subreddit communities. This depiction allows for the identification of trends, patterns, and possible correlations between stress levels and subreddit engagement throughout the years.

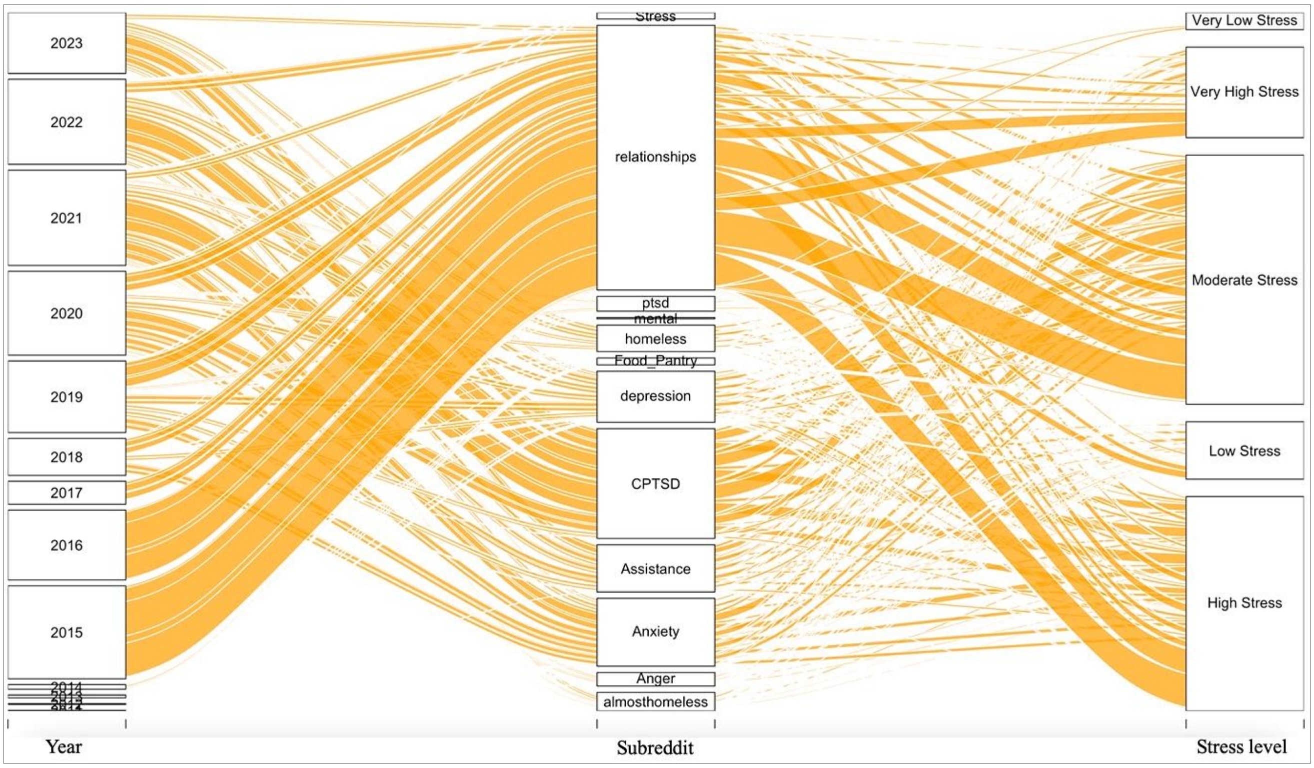


Figure 4. A temporal dynamics of stress levels across various subreddit categories over the years

Table 3. Descriptive statistic for stress and non-stress posts in each subreddit

Year	No. of Sub-Reddits	No. of posts in stress level									
		Very low		Low		Moderate		High		Very high	
		Mean (SD)	Max	Mean (SD)	Max	Mean (SD)	Max	Mean (SD)	Max	Mean (SD)	Max
2011	1	2 (NA)	2	2 (NA)	2	23 (NA)	23	2 (NA)	2	2 (NA)	2
2012	5	0 (0)	0	0 (3.49)	8	3 (5.89)	12	0 (3.49)	8	0 (0)	0
2013	5	1 (3.21)	8	0 (4.6)	10	7 (21.88)	53	0 (4.6)	10	1 (3.21)	8
2014	8	0.5 (3.4)	10	1 (8.59)	25	4 (44.72)	129	1 (8.59)	25	0.5 (3.4)	10
2015	9	0 (56.92)	171	2 (192.79)	581	6 (911.54)	2743	2 (192.79)	581	0 (56.92)	171
2016	8	0 (46.92)	133	0.5 (142.23)	403	6 (760.05)	2155	0.5 (142.23)	403	0 (46.92)	133
2017	10	0 (10.23)	33	1 (29.45)	94	10.5 (176.09)	569	1 (29.45)	94	0 (10.23)	33
2018	12	2.5 (9.21)	30	10.5 (31.72)	99	21 (142.26)	420	10.5 (31.72)	99	2.5 (9.21)	30
2019	12	8.5 (15.96)	53	28 (53.67)	172	94 (215.63)	762	28 (53.67)	172	8.5 (15.96)	53
2020	12	12 (16.15)	42	29 (50.6)	133	125.5(217.07)	642	29 (50.6)	133	12 (16.15)	42
2021	11	13 (27.13)	92	40 (83.9)	275	158 (334.76)	1162	40 (83.9)	275	13 (27.13)	92
2022	12	9.5 (18.73)	67	38.5 (68.89)	252	188.5(264.95)	963	38.5 (68.89)	252	9.5 (18.73)	67
2023	11	7 (12.59)	43	37 (41.86)	149	155 (203.35)	677	37 (41.86)	149	7 (12.59)	43

3.4. RQ4. Are there any correlations between polarity scores and the engagement metrics (Redditors’ scores, number of comments) in each stress level or not?

The analysis revealed significant correlations between the polarity scores and the Redditors’ scores, as assessed by Redditors across the total posts within the stress group. Additionally, these correlations were observed within three distinct stress levels: *very high*, *high*, and *moderate*. The Spearman’s rank correlation coefficients were found to be 0.037, 0.058, 0.017, and 0.034, respectively, for the aforementioned stress levels (all with $p < 0.05$). Likewise, the polarity scores exhibited correlations with the number of comments across the total posts within the stress group and the same three stress levels: *very high*, *moderate*, and *low*, with respective values of 0.036, 0.042, 0.032, and 0.045 (all with $p < 0.01$).

Figure 5 represents these findings visually, illustrating the relationship between polarity scores and Redditors' scores or the number of comments within different stress levels. The observed correlations further underscore the interconnectedness of sentiment expression and user engagement metrics in the context of stress-related discussions.

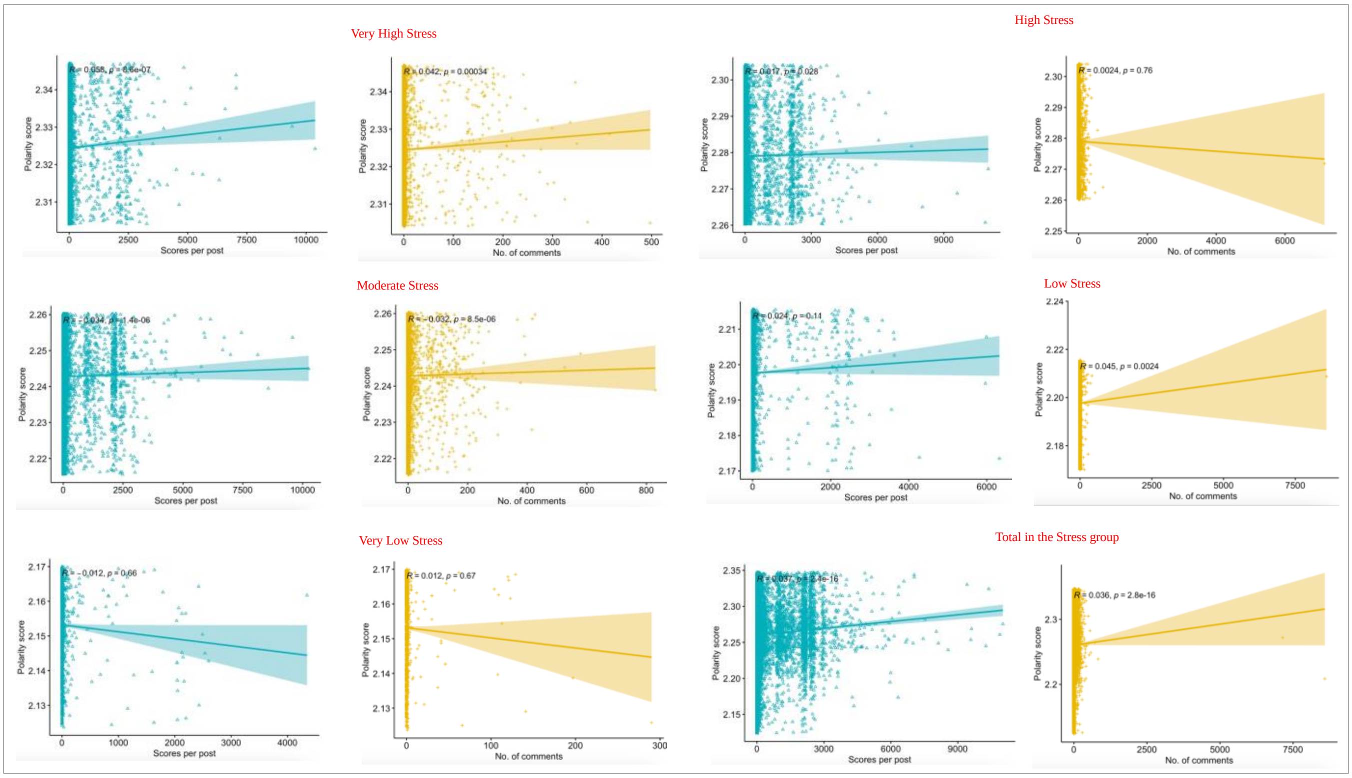


Figure 5. Correlations between polarity score and engagement metrics (Redditors' score, number of comments) in each stress level

The factors associated with polarity scores in the univariable models included (1) *Redditors' score*, (2) *number of comments*, (3) *total received awards*, (4) *golds*, and (5) *Up ratio*. We employed five univariable models, both before and after January 2020, among which the three models, namely *Redditors' score, number of comments*, and *total received awards*, yielded significant statistical results with p-values < 0.001. In the final multi-variable models, the significance patterns varied based on the time frame. Before January 2020, two variables—*Redditors' score* and *total received awards*—remained significant with p-values < 0.001. After January 2020, however, only one variable—*number of comments*—retained its significance with p-values < 0.001 (See Table 4).

Table 4. Linear models that investigate factors associated with polarity scores split by date (before and after January 2020)

	Univariable				Multi-variables			
	Date < January 2020		Date ≥ January 2020		Date < January 2020 [i]		Date ≥ January 2020 [ii]	
	β	p-value	β	p-value	β	p-value	β	p-value
Redditors' score	2.89E-06	1.32E-10	3.33E-06	3.35E-07	5.03E-06	2.51E-12	2.31E-06	0.011984
No. of comments	-9.959E-07	0.778	6.11E-05	9.30E-11			4.01E-05	0.000785
Total received awards	-0.007956	9.51E-06	-0.00029	2.13E-02	-7.12E-03	6.90E-05	-1.44E-04	0.263313
Golds	-0.0010513	0.795	-0.00129	0.593	5.03E-06	2.51E-12		
Up ratio	0.0016865	0.402	-0.00045	0.726				

(i) Intercept = 2.2528; SE = 0.001036; (ii) Intercept = 2.2569; SE = 0.00087

3.5. RQ5. Can we leverage NLP-driven techniques to find important themes?

RQ5.1. What is the theme based on the highly frequent words for each subreddit?

Figure 6 represents the top five words (bi-grams) extracted from stress-related posts of each subreddit, before and after January 2020. Each of these words is associated with a significance level, indicated by a p-value < 0.2. Particularly interesting is the term *result_ban*, identified within the *Assistance* subreddit (abbreviated as “as”), which attained the highest keyness score exceeding 235 in the period after January 2020. During the period, a substantial number of words originating from the *Assistance* and *relationships* subreddits exhibited robust keyness scores, surpassing the threshold of 50. In contrast, the *ptsd* subreddit (“pt”) predominantly displayed words characterized by significantly lowest keyness scores below 5. Before January 2020, the most prominent keyness score of 86 was identified within the *CPTSD* subreddit (“cp”). Conversely, words extracted from the *Anger* subreddit (“an”) yielded comparatively the lowest keyness scores within this timeframe. Notably, the computation of the keyness score is rooted in the chi-square (χ^2) statistic.

After January 2020, the terms *race_gender*, *social_media*, *wear_mask*, *result_ban*, *crisi_resourc*, *accept_risk*, and *manag_stress* exhibited high keyness scores in stress-related posts in their respective subreddits (almosthomeless, Anger, Anxiety, Assistance, CPTSD, homeless, and Stress) after January 2020. These keyness scores indicated active discussions around these terms within the specific topics of each subreddit during this period, reflecting emerging themes. The frequencies of these terms varied significantly, with minimum values of 10 and maximum values of 1,092, all accompanied by statistically significant p-values, most < 0.05. Additionally, other notable terms, such as *enjoy_life* in the *depression* subreddit, *follow_rule* in *Food_Pantry*, *eye_contact* in *mental health*, *diagnos_ptsd* in *ptsd*, and *white_lie* in *relationships* also demonstrated high keyness scores. These terms were actively discussed in stress-related posts during this period, with some experiencing an increase in frequency compared to the previous period (n_reference values were not equal to 0, and associated p-values were < 0.09). Conversely, several terms, such as *homeless_person*, *wrong_forgiv*, *phone_call*, *dirti_tree*, *perpetr_preoccupi*, *leav_hous*, varieti*_pack*, *gutter_punk*, *video_game*, *weed_trip*, and *nervous_system*, which were actively discussed before January 2020, saw reduced engagement during the later period. Their n_target values either decreased or became zero, with most of these terms exhibiting p-values < 0.05 (Figure 6).

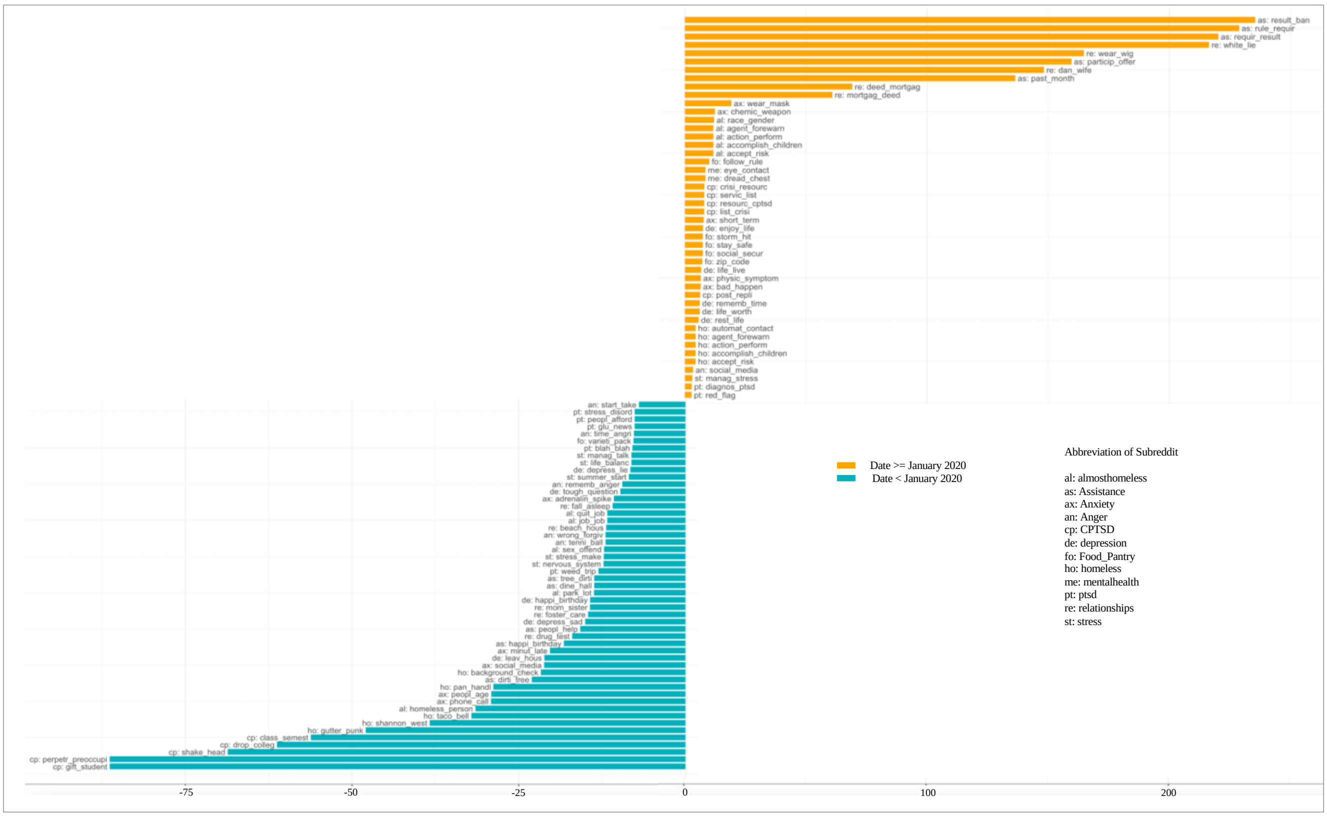


Figure 6. Top common words (bi-gram) in the stress-related posts between before and after January 2020

From these commonly mentioned terms, we identified themes for each subreddit before and after January 2020, shedding light on the various discussions within each community. For example, the *almosthomeless* subreddit focused on *Challenges & concerns of homeless individuals* and *Empowerment & risk management*, while *Anger* delved into *Managing anger & forgiveness* and *Digital anger & mental health*. Each subreddit had its own set of themes reflecting the evolving discourse in these stress-related discussions (Table 5).

Table 5. Themes of each subreddit from stress-related posts before and after January 2020

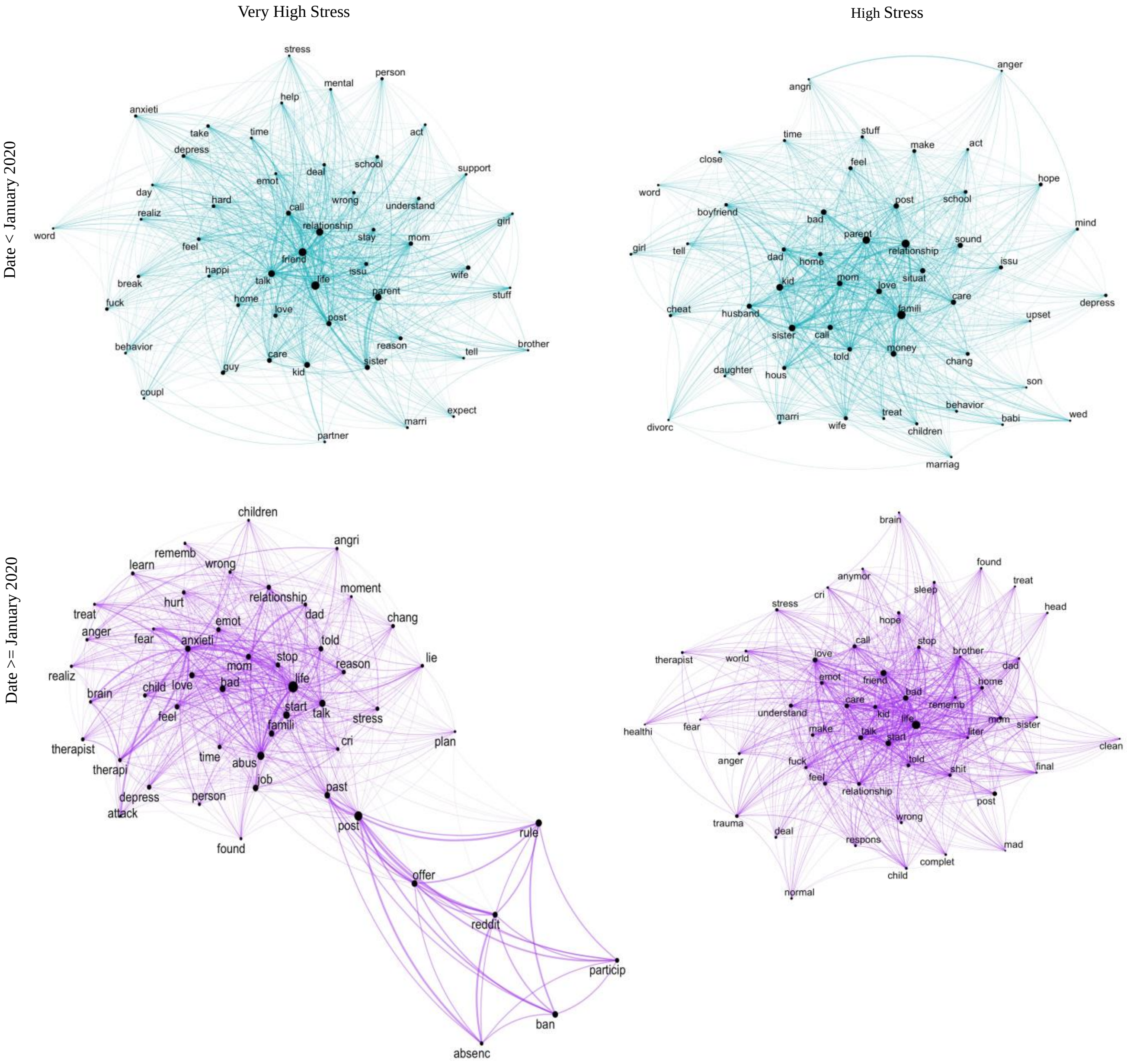
Date < January 2020	**Date ≥ January 2020**
Subreddit: ***Theme of subreddit*** words (n_target; n_reference)	*Subreddit*: ***Theme of subreddit*** words (n_target; n_reference)
almosthomeless: ***Challenges & concerns of homeless individuals*** homeless_person (0; 11)***; park_lot (6; 11)***; sex_offend (0; 5)***; quit_job (1; 6)***; job_job (1; 6)***	*almosthomeless:* ***Empowerment & risk management*** race_gender (39; 0)***; accept_risk (38; 0)***; accomplish_children (38;0)***; action_perform(38;0)*** agent_forewarn (38; 0)***
Anger: ***Managing anger & forgiveness*** wrong_forgiv (0;6)***; tenni_ball (0;6)***; rememb_anger (0;5)**; time_angri (6;10)**; start_take (0;4)**	*Anger:* ***Digital anger & mental health*** social_media (12;0)$^{\cdot}$; multipl_time (9; 0); mental_health (22;4); video_game (22;6); anger_angri (8;0)
Anxiety: ***Social anxiety & everyday stressors*** phone_call (12;32)***; peopl_age (1;17)***; social_media (15;30)***; minut_late (0;12)***; adrenalin_spike (0;7)***	*Anxiety:* ***Pandemic anxiety & health concerns*** wear_mask (37;0)***; chemic_weapon (24;0)***; short_term (15;0)***; physic_symptom (43;9)**; bad_happen (17;1)**
Assistance: ***Acts of kindness & assistance in everyday life*** dirti_tree (0;6)***; happi_birthday (1;6)****;* peopl_help (3;7)***; tree_dirti (0;4)***; dine_hall (0;4)***	*Assistance:* ***Rule changes & community participation*** result_ban (1,092;0)***; rule_requir (1,062;0)***; requir_result (1,023;0)***; particip_offer (746;0)***; past_month (639;0)***
CPTSD: ***Challenges in pursuing education & coping with CPTSD*** perpetr_preoccupi (0;6)***; gift_student (0;6)*** shake_head (3;7)**; drop_colleg (4;7)**; class_semest (1;3)**	*CPTSD:* ***Crisis resources & supportive services for CPTSD*** crisi_resourc (139; 0)**; list_crisi (138;0)** resourc_cptsd (138;0)**; servic_list (138;0)**; post_repli (142;1)**
depression: ***Coping with isolation & emotional struggles*** leav_hous (2;18)***; depress_sad (1;12)***; happi_birthday (0;11)***; tough_question (0;8)***; depress_lie (0;7)***	*depression:* ***Finding hope & meaning in life amid depression*** enjoy_life (15;1)**; life_live (17;2)** life_worth (16;2); rememb_time (16;2)**; rest_life (11;0)
Food_Pantry: ***Food assistance & donations*** varieti_pack (0;21)***; ounc_pack (0;13).; groceri_store (1;17)$_{\cdot}$; gift_card (4;27)$_{\cdot}$; food_pantri (5;28)$_{\cdot}$	*Food_Pantry:* ***Emergency assistance and safety measures*** follow_rule (5;0)**; social_secur (4;0)**; stay_safe (4;0)**; storm_hit (4;0)**; zip_code (10;7)**
homeless: ***Street survival & survival strategies*** gutter_punk (0;6)***; shannon_west (0;5)***; taco_bell (4;7)***; pan_handl (0;4)***; background_check (1;4)***	*homeless:* ***Coping strategies & support*** accept_risk (53;0)*; accomplish_children (53;0)*; action_perform (53;0)*; agent_forewarn (53;0)*; automat_contact (53;0)*
mental health: ***Coping mechanisms & seeking answers*** video_game (0;4); peopl_act (0;4); health_issu (0;4); dog_bark (0;4); answer_question (0;4)	*mental health:* ***Academic stress & anxiety*** eye_contact (2;0)***; dread_chest (2;0)***; 1st_semest (1;0); 2nd_sem (1;0); 3-4_hour (1;0)
ptsd: ***PTSD & various topics*** weed_trip (0;6)***; stress_disord (0;4)**; peopl_afford (0;4)**; glu_news (0;4)**; titl_ix (0;3)*	*ptsd:* ***PTSD awareness & understanding*** diagnos_ptsd (18;2).; red_flag (11;0).; sexual_assault (20;3).; risk_factor (9;0); peopl_understand (8;0)
relationships}: ***Family, challenges, & relationships*** drug_test (0;70)***; foster_care (1;69)***; mom_sister (0;59)***; beach_hous(0;49)***; fall_asleep (0;45)***	*relationships:* ***Relationship dynamics & intricacies*** white_lie (54;1)***; wear_wig (40;0)***; dan_wife (36;0)***; deed_mortgag (18;0)***; mortgag_deed (16;0)***
Stress: ***Stress management & lifestyle balance*** nervous_system (1; 5)***; stress_make (0; 4)***; summer_start (1;4)**; manag_talk (0;3)**; life_balanc (0;3)**	*Stress:* ***Coping with stress & mental health*** manag_stress (18;0)$^{\cdot}$; fatigu_syndrom (10;0); hard_time (9;0); stress_level (9;0); depress_anxieti (8;0)

Note: ***p < 0.001, **p < 0.01; *p < 0.05; $^{\cdot}$: < 0.1

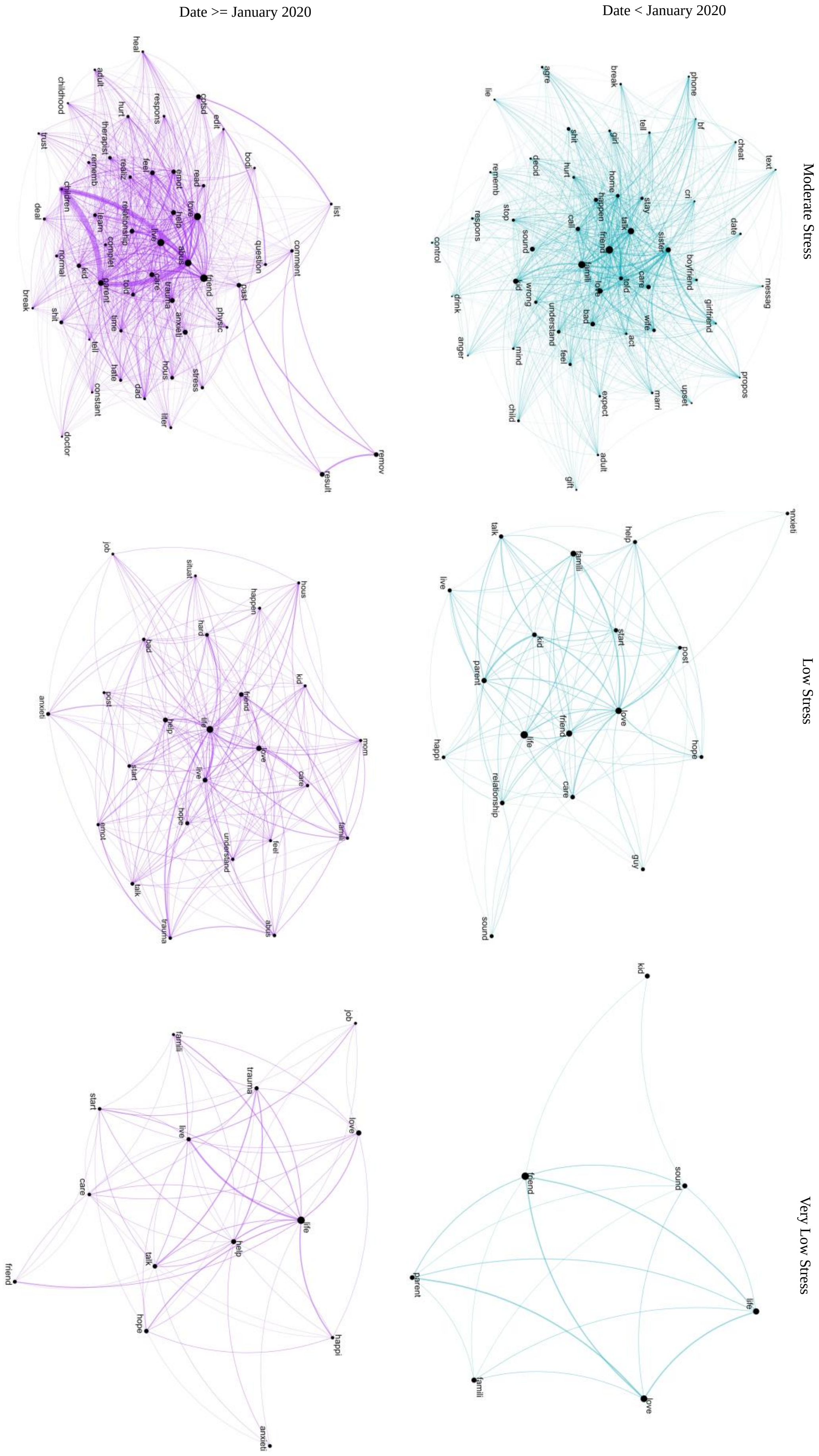
RQ5.2. What is the theme based on semantic networks for each stress level?

Semantic networks, which are based on word co-occurrence, were generated for five stress levels before and after January 2020, as depicted in Figure 7. Notably, the semantic networks representing the *very high*, *high*, and *moderate* stress levels featured the most frequently occurring keywords in across SubReddit discussions related to stress.

For the *very high-stress* category before January 2020, five keywords exhibited high frequencies, comprising *friend*, *life*, *relationship*, *parent*, and *talk*. However, after January 2020, the high-frequency keywords, including the terms *life*, *abus, bad*, *talk*, *famili*, *anxiety*, *job*, *abus*, *post*, *ban*, and *rule,* underwent a shift. In the *high-stress* category before January 2020, the high-frequency keywords were *relationship*, *famili*, *parent*, *kid*, *sister*, *love*, and *money*. Following January 2020, the high-frequency keywords shifted to *life*, *love, told, liter, call, care*, *relationship*, *start, talk*, and *understand*. Before January 2020, the *moderate-stress* category featured high-frequency keywords *famili, friend, talk, love, sister, care, home, call,* and *bad*. However, after January 2020, the high-frequency keywords transformed into *friend, live, parent, love, relationship, care, trauma, abus*, and *anxieti*. In the *low-stress* category before January 2020, high-frequency keywords included *famili*, *friend*, *life*, and *love*. Conversely, after January 2020, the high-frequency keywords shifted to *life, love, live*, *help*, *friend, hope, anxieti, emot, feel*, and *understand*. Lastly, in the *very low-stress* category, both before and after January 2020, the high-frequency keywords remained consistent with *friend*, *life*, and *love*. New high-frequency keywords, including *life*, *love*, *help*, *trauma*, and *help*, emerged after January 2020.



(a) Semantic networks of very high- and high-stress levels



(b) Semantic networks of moderate-, low-, and very low-stress levels

Figure 7. Semantic networks of each stress level before and after January 2020

These high-frequency words and their co-occurrence with other words allowed us to identify themes for each stress level before and after January 2020, ranging from *Coping with relationship stress* and *Coping with life challenges and mental health* for the *very high-stress* level to *Friendship and positive relationships* and *Coping with life's challenges and seeking support* for the *very low-stress* level (Table 6).

Table 6. Themes of each stress level from stress-related posts before and after January 2020

Date < January 2020	**Date ≥ January 2020**
Stress level: ***Theme of stress level*** Key elements: details of words	*Stress level*: ***Theme of stress level*** Key elements: details of words
Very High: ***Coping with relationship stress*** Friendships: friend, talk, relationship Relationships: relationship, talk, parent, love, care Communication: talk, call Support systems: call, home, care	*Very High:* ***Coping with life challenges & mental health*** Life challenges: life, famili, start, reason Anxiety & mental health: anxiety, fear, therapi, attack Abuse & family: abus, famili, child, treat Emotional well-being: bad, life, love, feel, famili Communication & support: talk, start, stress, famili, love Work & Job-Related Stress: job, life, abus, anxiety Online community rules: post, ban, rule
High: ***Family relationships & financial stress*** Family relationships: relationship, parent, famili, mom Parenting: parent, home, school, bad Siblings: sister, husband, kid Financial stress: money, hous, post,call,bad, situat	*High:* ***Emotional well-being & relationships*** Emotional well-being: life, love, told, liter, call, care, relationship Love & relationships: love, world, relationship, friend, bad Communication: told, liter, life, remember, mom, dad Mental health: call, brain, cri, stress, emot, sleep, found, treat Caring & understanding: care, understand, kid, feel, relationship Relationship dynamics: relationship, life, love, told, liter, call
Moderate: ***Relationships & family in moderate stress discussions*** Family bonds: famili, love, friend, happen, call, kid Friendships: friend, talk, home, told, love, sister Communication: talk, friend, home, stay, sister, care Mental health: call, brain, cri, stress, emot, sleep, found	*Moderate:* ***Coping with trauma, relationships, Emotional support*** Supportive friendships: friend, abus, love, live, care, help, trauma Life & relationships: live, care, relationship, emot, love Parent-child dynamics: parent, children, emot, respons, learn, hate Love & relationships: love, famili, wife, kid, understand, friend Relationship dynamics: relationship, live, complet, learn, friend Caring & emotional well-being: care, abus, complet, love
Low: ***Positive relationships & emotional support*** Family bonds: famili, love, life, live, parent, help Supportive friendships: friend, life, love, parent Embracing life: live, famili, love, friend, life, help, care Love & caring: love, care, famili, kid, hope, post, help	*Low:* ***Coping & emotional support*** Embracing life: life, live, friend, love, help, bad, hard, famili Love & relationships: love, care, friend, life, live, feel, kid Active living: live, life, understand, help, hope, start, love, hous Seeking help: help, life, friend, hard, hope, trauma, understand
Very Low: ***Friendship & positive relationships*** Friendship bonds: friend, life, love, parent Embracing life: life, love, friend, famili Love & affection: love, life, friend, parent	*Very Low:* ***Coping with life's challenges & seeking support*** Navigating life: life, love, live, happi, famili, friend Emotional support: love, life, trauma, care Seeking Help: help, trauma, start, life, hope Coping with trauma: trauma, talk, help, start, life

4. Discussions

This study employed Subreddit data to investigate Redditors' mental stress before and after January 2020, as evidenced by their activities on social media. We delved into and characterized how Redditors' actions related to posting and commenting and their emotional expressions evolved. The results indicated that the level of Redditors' engagement in subreddits, their posting and commenting frequency, and their emotional expressions were associated with mental stress. We provide a detailed discussion of the main findings.

Analyzing Redditors' engagement and the popularity of content in subreddits using metrics such as Redditors' scores and the number of comments is a well-established method in Reddit analytics, providing valuable insights into the dynamics of online communities. Posting frequency and emotional expression from Redditors in subreddits were influenced by significant pandemic events [29, 30]. Our research considered the relationship between Redditors' scores and the number of comments for each subreddit, enabling us to identify distinct engagement patterns. Notably, the presence of high positive correlations suggests a strong interest in and active participation in subreddit discussions. These findings underline the vibrant and responsive nature of Reddit communities, which adapt and engage in response to various events and topics (see Figure A1 in Appendix I).

Sentiment analysis plays a crucial role in illustrating the stress levels of posts for the public. We can gain a deeper understanding of the stress levels expressed in the community by assessing the emotional tone of posts. Applying machine learning methods to classify stress from social media had high accuracy [8–10, 31]. In this study, we deployed the three high-performance machine learning methods (TextBlob, VADER, and Flair), combined with a majority voting approach, to identify stress across various subreddits. The majority voting method [32] aims to reduce the impact of potential biases or inaccuracies in individual sentiment analysis models and enhance the robustness and reliability of stress level determination. This approach can lead to more accurate assessments of stress levels in posts within subreddits, improving the overall reliability of sentiment-based stress analysis. Our study found that there was a significant increase in negative sentiments via five stress levels from subreddit data after January 2020 (the COVID-19 pandemic) compared to before January 2020 (the pre-pandemic). This finding is consistent with recent studies that have linked negative sentiments [6, 31]. This parallel increase in negative sentiments and stress levels underscores the pandemic's profound influence on public opinion and mental well-being, which manifested strongly in digital spaces (see Figure 4 and Table 3).

As shown by the results, among all the posts in the stress group, two factors, namely the number of comments and Redditors' scores, displayed significant statistical relationships with polarity scores (both with p-values < 0.001). When we delved deeper into the specific stress levels, we found that the majority (four out of five) of stress levels also displayed significant statistical associations between these two factors and polarity scores (with p-values < 0.05). Furthermore, our research revealed a noteworthy finding: among the five factors under study, the number of comments displayed statistically significant impacts on the polarity scores of stress-related posts after January 2020 (with all p-values < 0.001). This finding is consistent with recent studies that have highlighted the influence of significant events, such as the pandemic, on Redditors' posting frequency and emotional expression within subreddits [29, 30]. These findings collectively emphasize the importance of considering factors such as posting frequency and user engagement when assessing the sentiment and emotional dynamics of stress-related discussions on Reddit. Moreover, these findings also highlight the dynamic nature of factors influencing polarity scores over time and underscore the importance of considering various engagement metrics and user behaviors in understanding sentiment dynamics within stress-related discussions (see Figure 5 and Table 4).

NLP-driven techniques have proven to be powerful tools for identifying and extracting significant themes from data, which can be valuable for the public's understanding [12, 14]. Our study employed functions from the quanteda package, specifically designed for handling text data, to uncover themes within stress-related posts. Our approach aligns with similar methods in other studies utilizing the quanteda package for theme extraction [13, 33–38]. The current study considered the imbalance in the total number of posts across different subreddits and addressed this by identifying themes for each subreddit before and after January 2020. Moreover, we also analyzed themes across five stress levels before and after January 2020 via semantic networks (before January 2020, stress was related to relationships; meanwhile, after January 2020, stress-related life challenges, health, and care). These approaches allowed us to analyze themes comprehensively across various stress-related discussions on Reddit. Experts were involved to ensure the relevance and accuracy of the identified themes.

The study is subject to several limitations. Firstly, relying on Reddit data introduces potential biases, as Reddit users may not represent the broader population, limiting the generalizability of our findings to offline or other online communities. Additionally, the dataset's quality can vary, which may impact the accuracy of our analysis. Although NLP-driven techniques are potent tools, they need to capture the complete context of discussions. Our focus on changes before and after January 2020, during the COVID-19 pandemic, might only partially account for other significant factors in the selected period. While expert validation enhanced the precision of our identified themes, the inherent subjectivity of theme identification remains a concern. Finally, external events and potential publication bias within the Reddit platform may not have been comprehensively considered.

5. Conclusion

This study aimed to examine trends, sentiments, and topics within diverse subreddits dedicated to the online mental stress community on Reddit before and after January 2020. The findings revealed robust engagement patterns, indicating a strong interest in discussions within the subreddit when correlations between Redditors' scores and the number of comments were highly positive. Redditors openly shared experiences and expressed emotions related to living with mental stress. The analysis emphasized a significant presence of very high, high, and moderate stress levels in posts, reflecting the multifaceted challenges associated with the condition. Dynamic conversations on mental stress covered various aspects, including mood swings, diagnosis, medication, coping strategies, support, and stress management. These revelations underscore the importance of tailored treatment and support. The research framework demonstrated adaptability for providing health condition-related implications through social media analysis. In summary, this study enhances our understanding of how individuals communicate stress on social media, offering avenues for future exploration, including unmet information needs and aspects requiring professional attention, ensuring practical relevance and depth.

6. Declarations

6.1. Author Contributions

Conceptualization, T.T.A., L.D.T.T., and N.N.H.; methodology, T.T.A., L.D.T.T., A.T.D., and N.N.H.; software, N.N.H. and A.T.D.; validation, T.T.A., L.D.T.T., and N.N.H.; formal analysis, T.T.A. and A.T.D.; resources, N.N.H. and A.T.D.; data curation, T.T.A. and L.D.T.T.; writing—original draft preparation, T.T.A., L.D.T.T., A.T.D., and N.N.H.; writing—review and editing, T.T.A. and L.D.T.T.; visualization, A.T.D. and N.N.H.; supervision, T.T.A. and L.D.T.T. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research.

6.4. Acknowledgements

The authors would like to thank the School of Informatics and Informatics Innovation Center of Excellence (IICE), Walailak University, for supporting this research.

6.5. Declaration of Competing Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript.

7. References

[1] Young, C. C., & Dietrich, M. S. (2015). Stressful life events, worry, and rumination predict depressive and anxiety symptoms in young adolescents. Journal of Child and Adolescent Psychiatric Nursing, 28(1), 35-42. doi:10.1111/jcap.12102.

[2] Moore, C. M., & Chuang, L. M. L. (2017). Redditors revealed: Motivational factors of the Reddit community. Proceedings of the Annual Hawaii International Conference on System Sciences, Volumes: January, 2313–2322. doi:10.24251/hicss.2017.279.

[3] Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J. C., & Ungar, L. H. (2019). Understanding and measuring psychological stress using social media. Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019, 214–225. doi:10.1609/icwsm.v13i01.3223.

[4] Beyari, H. (2023). The Relationship between Social Media and the Increase in Mental Health Problems. International Journal of Environmental Research and Public Health, 20(3), 2383. doi:10.3390/ijerph20032383.

[5] Inamdar, S., Chapekar, R., Gite, S., & Pradhan, B. (2023). Machine Learning Driven Mental Stress Detection on Reddit Posts Using Natural Language Processing. Human-Centric Intelligent Systems, 3(2), 80–91. doi:10.1007/s44230-023-00020-8.

[6] Abro, H. U., Shah, Z. S., & Abbasi, H. (2022). Analysis Of COVID-19 Effects on Wellbeing - Study of Reddit Posts Using Natural Language Processing Techniques. 2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering, ETECTE 2022 - Proceedings, 1–7. doi:10.1109/ETECTE55893.2022.10007300.

[7] Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (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), 22635. doi:10.2196/22635.

[8] Saha, K., Kim, S. C., Reddy, M. D., Carter, A. J., Sharma, E., Haimson, O. L., & Choudhury, M. D. E. (2019). The language of LGBTQ+ minority stress experiences on social media. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–22. doi:10.1145/3359191.

[9] Febriansyah, M. R., Nicholas, Yunanda, R., & Suhartono, D. (2022). Stress detection system for social media users. Procedia Computer Science, 216, 672–681. doi:10.1016/j.procs.2022.12.183.

[10] Shen, J. H., & Rudzicz, F. (2017). Detecting Anxiety through Reddit. Computational Linguistics and Clinical Psychology, 58–65.

[11] Nayak, S., Mahapatra, D., Chatterjee, R., Parida, S., & Dash, S. R. (2022). A Machine Learning Approach to Analyze Mental Health from Reddit Posts. Smart Innovation, Systems and Technologies, 271, 357–366. doi:10.1007/978-981-16-8739-6_33.

[12] Naseem, S. S., Kumar, D., Parsa, M. S., & Golab, L. (2020). Text mining of COVID-19 discussions on reddit. Proceedings - 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2020, 687–691. doi:10.1109/WIIAT50758.2020.00104.

[13] Gao, S., Pandya, S., Agarwal, S., & Sedoc, J. (2021). Topic Modeling for Maternal Health Using Reddit. Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, 69–76.

[14] Stevens, H. R., Acic, I., & Rhea, S. (2021). Natural language processing insight into LGBTQ+ youth mental health during the COVID-19 Pandemic: Longitudinal content analysis of anxiety-provoking topics and trends in emotion in lgbteens microcommunity subreddit. JMIR Public Health and Surveillance, 7(8), 29029. doi:10.2196/29029.

[15] Zhu, J., Yalamanchi, N., Jin, R., Kenne, D. R., & Phan, N. H. (2023). Investigating COVID-19's Impact on Mental Health: Trend and Thematic Analysis of Reddit Users' Discourse. Journal of Medical Internet Research, 25, 46867. doi:10.2196/46867.

[16] Papakyriakopoulos, O., Engelmann, S., & Winecoff, A. (2023). Upvotes? Downvotes? No Votes? Understanding the relationship between reaction mechanisms and political discourse on Reddit. Conference on Human Factors in Computing Systems - Proceedings, 1–28. doi:10.1145/3544548.3580644.

[17] Rivera, I. (2019). RedditExtractoR: Reddit Data Extraction Toolkit. A collection of tools for extracting structured data from. Version: 3.0.9. RedditExtractoR archive. Available online: https://cran.r-project.org/web/packages/RedditExtractoR/index.html (accessed on March 2023).

[18] Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30), 774. doi:10.21105/joss.00774.

[19] Kassambara, A. (2023). ggpubr: 'ggplot2' Based Publication Ready Plots. Version: 0.6.0. ggpubr archive. Available online: https://cran.r-project.org/package=ggpubr (accessed on March 2024).

[20] Liu, B. (2020). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Second Edition. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Second Edition. Cambridge University Press, Cambridge, United Kingdom. doi:10.1017/9781108639286.

[21] Tran, T. A., Duangsuwan, J., & Wettayaprasit, W. (2021). Novel framework for aspect knowledge base generated automatically from social media using pattern rules. Computer Science, 22. doi:10.7494/csci.2021.22.4.4028.

[22] Tran, T. A., Duangsuwan, J., & Wettayaprasit, W. (2021). A new approach for extracting and scoring aspect using SentiWordNet. Indonesian Journal of Electrical Engineering and Computer Science, 22(3), 1731–1738. doi:10.11591/ijeecs.v22.i3.pp1731-1738.

[23] Loria, S. (2018). textblob Documentation. Release 0.15, 26 April, 1-73. Available online: https://readthedocs.org/projects/textblob/downloads/pdf/latest/ (accessed on March 2024).

[24] Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and Vader sentiment. Lecture Notes in Engineering and Computer Science, 2239, 12–16.

[25] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. NAACL 2019, Annual Conference of the North American Chapter of the Association for Computational Linguistics Demonstrations, 54–59.

[26] Fliege, H., Rose, M., Arck, P., Walter, O. B., Kocalevent, R. D., Weber, C., & Klapp, B. F. (2005). The Perceived Stress Questionnaire (PSQ) reconsidered: Validation and reference values from different clinical and healthy adult samples. Psychosomatic Medicine, 67(1), 78–88. doi:10.1097/01.psy.0000151491.80178.78.

[27] Bojanowski, M. (2016). Creating Alluvial Diagrams. The Comprehensive R Archive Network. Available online: https://cran.r-project.org/web/packages/alluvial/vignettes/alluvial.html (accessed on March 2024).

[28] R Programming Language. (2022). The R Project for Statistical Computing. Available online: https://www.r-project.org/about.html. (accessed on March 2024).

[29] Zhang, S., Liu, M., Li, Y., & Chung, J. E. (2021). Teens' social media engagement during the covid-19 pandemic: A time series examination of posting and emotion on reddit. International Journal of Environmental Research and Public Health, 18(19), 10079. doi:10.3390/ijerph181910079.

[30] Veselovsky, V., & Anderson, A. (2023). Reddit in the Time of COVID. Proceedings of the International AAAI Conference on Web and Social Media, 17, 878–889. doi:10.1609/icwsm.v17i1.22196.

[31] Yan, T., & Liu, F. (2022). COVID-19 sentiment analysis using college subreddit data. PLoS ONE, 17(11 November), 275862. doi:10.1371/journal.pone.0275862.

[32] Ismail, Q., Obeidat, R., Alissa, K., & Al-Sobh, E. (2022). Sentiment Analysis of COVID-19 Vaccination Responses from Twitter Using Ensemble Learning. 2022 13th International Conference on Information and Communication Systems, ICICS 2022, 321–327. doi:10.1109/ICICS55353.2022.9811132.

[33] Czymara, C. S., Langenkamp, A., & Cano, T. (2021). Cause for concerns: gender inequality in experiencing the COVID-19 lockdown in Germany. European Societies, 23(S1), S68–S81. doi:10.1080/14616696.2020.1808692.

[34] Lemay, D. J., Baek, C., & Doleck, T. (2021). Comparison of learning analytics and educational data mining: A topic modeling approach. Computers and Education: Artificial Intelligence, 2, 100016. doi:10.1016/j.caeai.2021.100016.

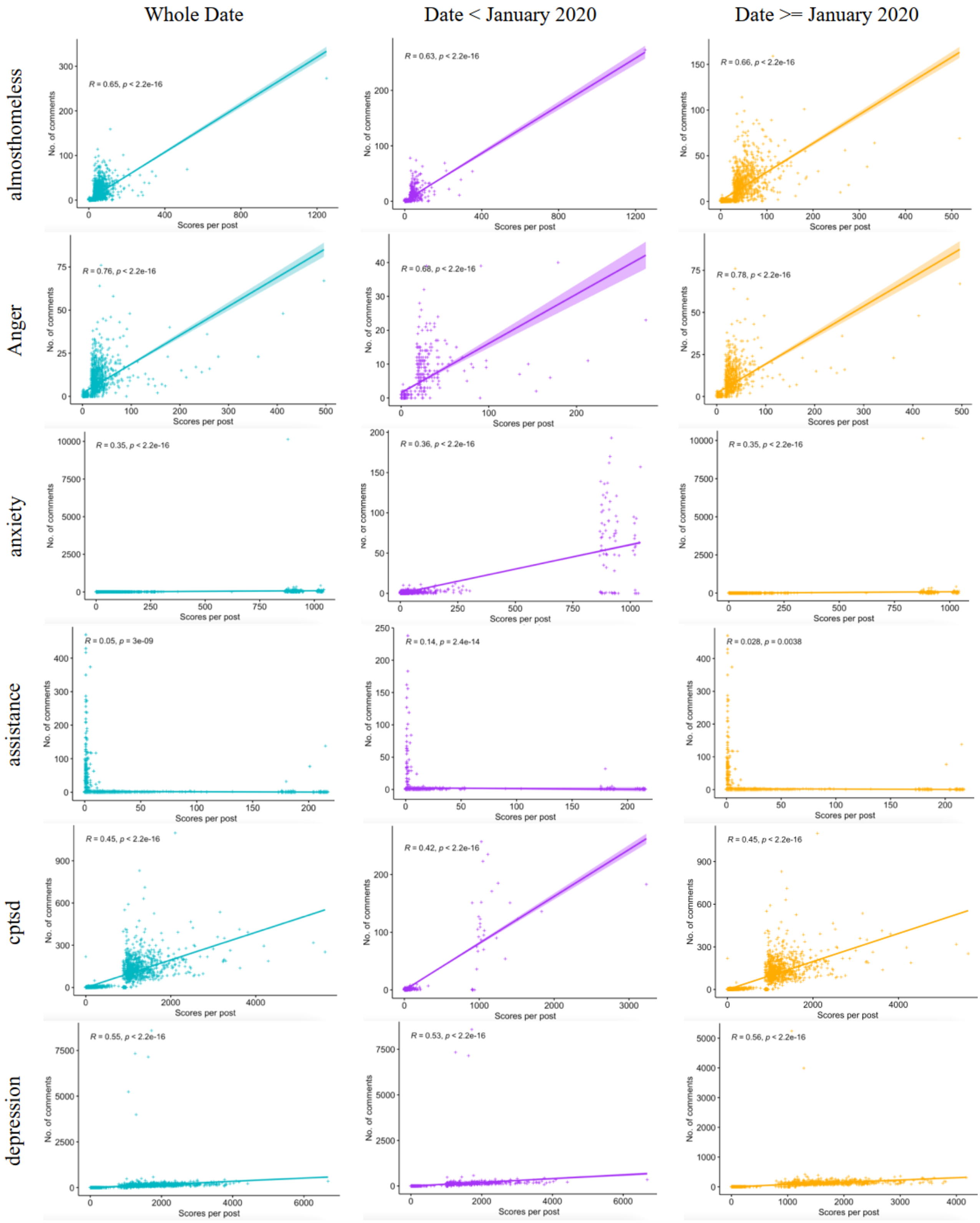
[35] Rosenberg, J. M., & Krist, C. (2021). Combining Machine Learning and Qualitative Methods to Elaborate Students' Ideas About the Generality of their Model-Based Explanations. Journal of Science Education and Technology, 30(2), 255–267. doi:10.1007/s10956-020-09862-4.

[36] Gaur, L., Jhanjhi, N. Z., Bakshi, S., & Gupta, P. (2022). Analyzing Consequences of Artificial Intelligence on Jobs using Topic Modeling and Keyword Extraction. Proceedings of 2nd International Conference on Innovative Practices in Technology and Management, ICIPTM 2022, 435–440. doi:10.1109/ICIPTM54933.2022.9754064.

[37] Yang, L. (2023). Mining and visualizing large-scale course reviews of LMOOCs learners through structural topic model. PLoS ONE, 18(5 May), 284463. doi:10.1371/journal.pone.0284463.

[38] Guerra, A. (2023). Sentiment analysis for measuring hope and fear from Reddit posts during the 2022 Russo-Ukrainian conflict. Frontiers in Artificial Intelligence, 6, 1163577. doi:10.3389/frai.2023.1163577.

Appendix I


Whole Date
Date < January 2020
Date >= January 2020
almosthomeless
Anger
anxiety
assistance
cptsd
depression
No. of comments
Scores per post
R = 0.65, p < 2.2e-16
R = 0.63, p < 2.2e-16
R = 0.66, p < 2.2e-16
R = 0.76, p < 2.2e-16
R = 0.68, p < 2.2e-16
R = 0.78, p < 2.2e-16
R = 0.35, p < 2.2e-16
R = 0.36, p < 2.2e-16
R = 0.35, p < 2.2e-16
R = 0.05, p = 3e-09
R = 0.14, p = 2.4e-14
R = 0.028, p = 0.0038
R = 0.45, p < 2.2e-16
R = 0.42, p < 2.2e-16
R = 0.45, p < 2.2e-16
R = 0.55, p < 2.2e-16
R = 0.53, p < 2.2e-16
R = 0.56, p < 2.2e-16

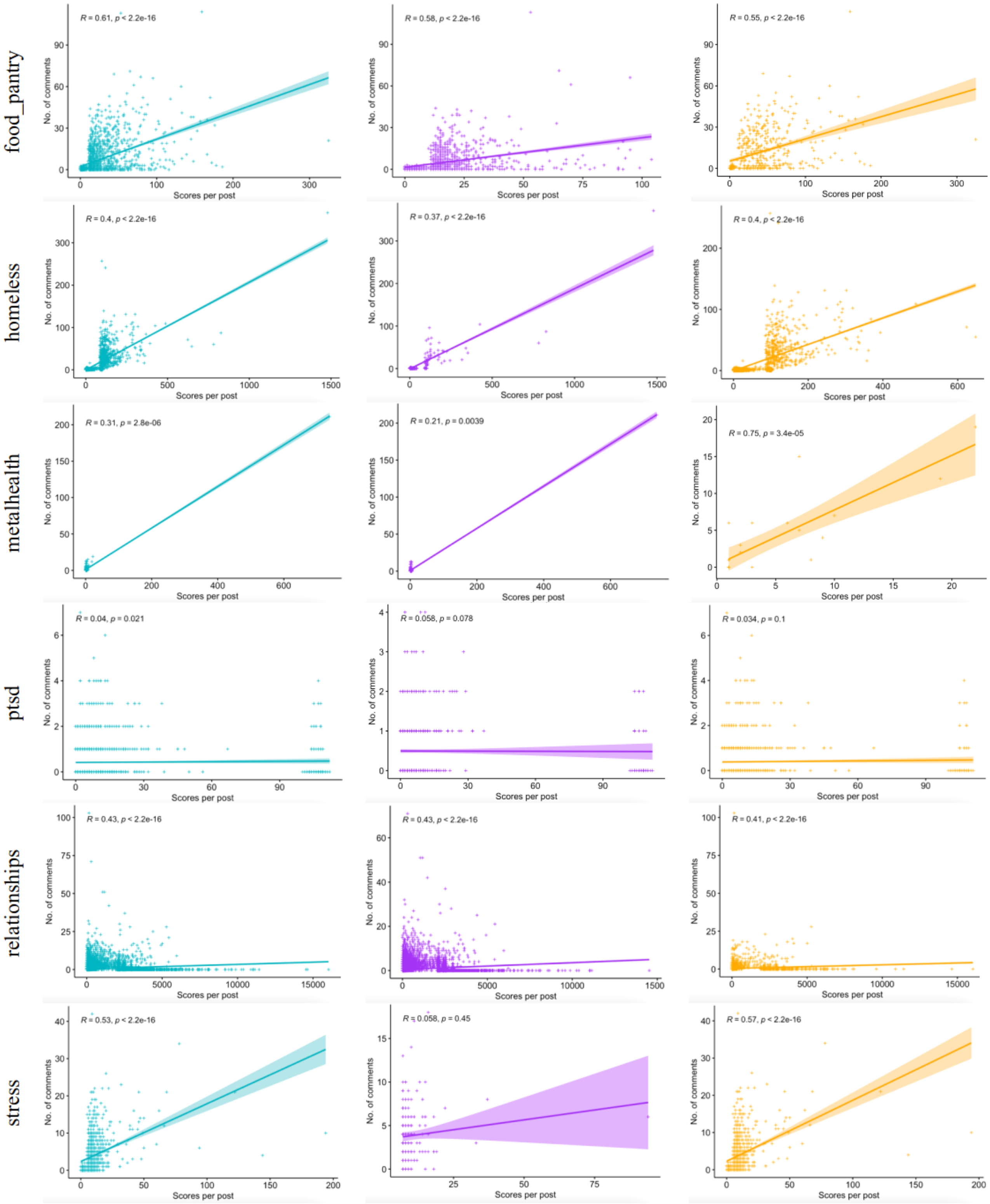


Figure A1. Correlations between users' scores and number of comments in each subreddit split by date: a whole (left), date < January 2020 (middle), and date >= January 2020 (right)