Group 8

Social media communities and mental well-being discussion: a data-driven exploration

Valentina Brivio, Edoardo Palombi, Mariapia Tedesco, Arianna Zottoli

Academic Year 2023-24

Table of Contents

1. Abstract

This research explores the impact of social media communities on individuals' mental well-being through a variegated data-driven analysis. Our study is motivated by the dual nature of the impact of social media on mental health, which could have both beneficial and harmful effects according to different theories.
The dataset was extracted from Reddit via the Reddit API, comprising 2847 posts from the main mental health-related subreddits. After rigorously preprocessing the data, topic modeling analysis identified the following predominant themes within the community discussions: depression, anxiety, anger, stress, and relationship issues.
Afterward, we collected posts from subreddits on these relevant topics and conducted a sentiment-emotion analysis. The first step was an exploratory analysis, which revealed, as expected, a generally negative sentiment across the posts, with emotions such as sadness and anger predominating, except for anxiety-related subreddits where posts with positive sentiment and emotions are in a quite high percentage. Once we extracted comments to posts (from other users) and the replies of posts' authors to these comments, the second step was a comparative analysis, in which we computed the average change in sentiment from posts to answers. We found that, overall, all types of comments seem to be followed by a positive sentiment change.
Finally, an important aspect of our analysis involved an interactive survey, engaging 215 respondents to understand the influence of social media engagement and interaction with mental health-related content on their mental health. We performed a controlled experiment and analyzed the effect that two different types of comments may have on an individual's emotional wellness.
Despite the presence of potential stressors, this study highlights the positive influence of interaction within online communities on mental wellbeing, serving as sources of release, support and sharing.

2. The problem

In recent years, the influence of social media on daily life has extended beyond mere digital interaction, fundamentally altering the landscape of social bonds and community building. This phenomenon provides the motivation to examine their impact on psychological well-being.

The academic consensus on the overall impact of social media use remains divided. According to the *"Displaced Behavior Theory"*[1], the link between social media and mental-health issues may stem from the reduction in face-to-face interactions. Engaging in sedentary activities like social media use diminishes the time available for in-person socializing, which is traditionally protective against mental disorders. In contrast, social media offers continuous opportunities for connectivity and interaction, regardless of time or location. This readily available communication can be particularly valuable for enabling social interaction among individuals with mental disorders who find face-to-face interactions challenging[2].

[1] It states that the more time humans spend on social media, the less time they spend in the real world.
[2] Naslund, J.A., Bondre, A., Torous, J. et al. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. J. technol. behav. sci. 5, 245–257 (2020). https://doi.org/10.1007/s41347-020-00134-x

The hypothesis driving this research posits that a sense of belonging to digital groups could potentially mirror the mental-health benefits traditionally associated with physical community engagement, especially for individuals with mental disorders[3].

The significance of this study is emphasized by the growing global mental-health crisis, with the *World Health Organization* reporting that approximately 280 million people worldwide suffer from depression[4]. Despite these alarming statistics, access to professional mental-health care remains limited. Mental disorders in low- and middle-income countries (LAMIC) often fail to attract global health policy attention[5], leading many individuals to turn to online communities for support.

From these premises, we argue that thoughtful engagement with social media can actually improve mental well-being and maybe ameliorate the conditions of those suffering from mental disorders (such depression and anxiety), hopefully making up (at least partially) for the lack of affordable social care around the world.

3. Data acquisition

We extracted the data from Reddit, an online platform hosting a wide variety of user-generated content such as texts, links, images, videos, and more, organized into communities called "subreddits". To this end, we used the Reddit API (Application Programming Interface) via OAuth (Open Authorization), an open standard for authorization that allows users to grant third parties limited access to their resources on a web service (such as Reddit) without sharing their credentials (such as passwords). Consequently, we created a web app as developers, containing a name, a description, a redirect URI pointing to a URL on a web server under our control, a public Client ID, and a secret Client ID.
To make requests to the Reddit API via OAuth, we needed to acquire an authorization token on behalf of a user, letting “reddit.com” know that we allowed the app to connect to our account.
For this reason, we sent the following URL to our account via our app:

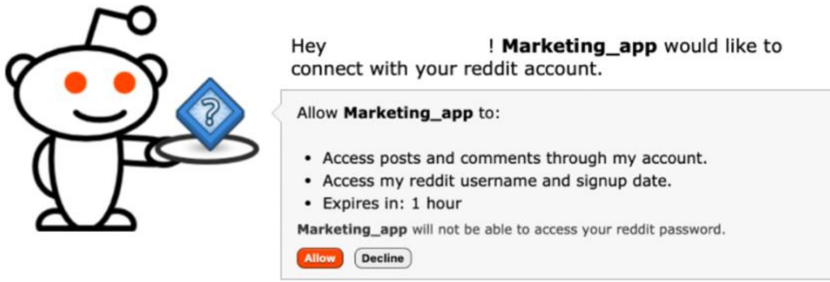
“https://www.reddit.com/api/v1/authorize?client_id=CLIENT_ID&response_type=TYPE&state=RANDOM_STRING&redirect_uri=URI&duration=DURATION&scope=SCOPE_STRING”

where:

- *client_id* is the Client ID generated during the app registration;
- *response_type* is the type of the response expected, commonly set to “*code*” for obtaining an authorization code;
- *state* is a unique, possibly random string for each authorization request;
- *redirect_uri* is the redirect URI specified during the app registration;
- *duration* specifies the duration of the authorization, that we set as “*temporary*”;

[3] Palis, H., Marchand, K., & Oviedo-Joekes, E. (2020). The relationship between sense of community belonging and self-rated mental health among Canadians with mental or substance use disorders. Journal of Mental Health, 29(2), 168–175. https://doi.org/10.1080/09638237.2018.1437602
[4] https://www.who.int/news-room/fact-sheets/detail/depression
[5] Patel V. Mental health in low- and middle-income countries. Br Med Bull. 2007;81-82:81-96. doi: 10.1093/bmb/ldm010. Epub 2007 Apr 30. PMID: 17470476.

- *scope* determines what actions the application can perform on behalf of the user. To achieve our goal, we used the scope "*read*" to access posts and comments through our account and the scope "*identity*", to access the reddit username and signup date.



Once we allowed our application, the browser redirected to our app's registered redirect URI, containing the authorization code. Subsequently, we made a post request with this code to the API endpoint URL "*https://www.reddit.com/api/v1/access_token*" to retrieve our access token.
After obtaining the access token, we collected the title, text, and authors of the top posts of the year from the three main subreddits on mental health: "*r/mentalhealth*" (459K members), "*r/mentalillness*" (151K members) and "*r/MentalHealthSupport*" (44K members).
The final dataset contains a total of 2847 posts.

4. Topic modelling

Our research started with a topic modeling analysis on posts collected from the three mental health-related subreddits.
This initial exploratory phase aims to identify the predominant themes discussed within these online communities. Our dataset includes nearly 3,000 posts, each averaging 80 characters in length, comfortably surpassing the literature's recommended minimum of 20 characters per document for such analyses.

To prepare the data for analysis, we first performed text normalization on the dataset, which involved both *stemming* and *lemmatizing* the words to reduce them to their base or dictionary forms. We also removed *stopwords* and punctuation to further cleanse the data. The cleaned texts, comprising both the titles and the contents of the posts, were then *tokenized* into words. These tokens were assembled into lists, creating a structured representation of our documents.

Following tokenization, we constructed a dictionary that only included *bigrams* occurring at least three times across the documents. Each unique token was assigned a distinct identifier. This dictionary was refined by excluding tokens that appeared in fewer than 100 documents or in more than 60% of the documents, ensuring the retention of only the most relevant words given the diverse and informal nature of social media language.

The corpus of texts was then converted into a *Term Frequency-Inverse Document Frequency* ***(TF-IDF)*** format, highlighting the significance of each word relative to the document set. With this TF-IDF corpus, we trained a *Latent Dirichlet Allocation* ***(LDA)*** model across a spectrum of topic counts, ranging from 2 to 10. To evaluate the model's effectiveness, we employed two coherence metrics: ***UMass*** and ***C_V***. The UMass coherence score, which ranges from -2 to 0, assesses topic quality based on word co-occurrence within documents. Meanwhile, the C_V measure, which ranges from 0 to 1, provides a broader evaluation of topic coherence.

Despite the challenges posed by the informal and varied language typical of social media, our meticulous approach to data cleaning yielded the most effective results, even though coherence scores remained modest. The highest scores obtained were -1.05 for UMass and 0.28 for C_V. These indexes guided our selection of the optimal number of topics, balancing both metrics to maximize interpretability.

The final LDA model was trained on the TF-IDF corpus with specific settings to enhance stability and interpretability: the model updates every 200 instances and uses a random to ensure reproducibility. Key parameters included an *alpha* setting of 0.5 to assume a moderate level of topic distribution across documents, along with adjustments to the *chunksize* and number of passes to influence convergence and topic detail. This structured approach allowed us to derive a set of topics that are both stable and meaningful, providing insightful themes prevalent among mental-health discussions on social media. The result of such model provides 5 topics, each composed of a series of words with different contributions (*Exhibit* 1)

Topic	**Title**	**Top 5 relevant terms**	**% of tokens**
Topic 1	Mental Health Management	mental, health, suicide, ill, help	18.9 %
Topic 2	Depression and Anxiety	depression, anxiety, sleep, therapy, medication	18.2%
Topic 3	Stress in daily life	work, job, school, lose, week	21.0%
Topic 4	Anger and Frustration	f*ck, look, die, guy, happen	21.3%
Topic 5	Expressing Emotions and Concerns	tell, friend, talk, cry, hurt	20.7%

Exhibit 1: The 5 topics identified through topic modelling and relevant tokens.

Based on the contributions of different words to each topic (*Appendix* C) and most representative comments (*Appendix C.1*), the following titles have been defined to represent the major themes discussed:

1) **Mental-health Management**: This topic primarily navigates serious discussions, often inquisitive, about mental-health statuses, among which trauma, suicide and critical struggles to individuals.
2) **Depression and Anxiety**: This discussion centers around the mental disorders of Depression and Anxiety, but mainly around the solutions undertaken to manage such conditions, among which therapy or medications.
3) **Stress in Daily Life**: Conversations in this category often relate to the pressures of work, education, and personal relationships. Key words like "lose," "year," "week," "day," "new," "home," and "life" suggest themes of transition, and the struggle to balance daily responsibilities with long-term goals.
4) **Anger and Frustration**: This topic captures discussions about feelings of anger, frustration, and generally negative emotions towards life and societal issues.

5) **Expressing Emotions and Concerns**: This topic explores how individuals communicate their struggles and emotions with close friends and family, often highlighting feelings of loneliness and invisibility.

From the intertopic distance map (*Exhibit* 2), there appears to be a partial overlap between Topics 4 and 5, suggesting that they may share several terms and potentially represent different facets of a similar issue. It appears in fact very plausible that expressing emotions and feelings related to mental health may often be accompanied by frustrations and anger. Similarly, Topics 2 and 3 appear close to one another; the discussions around challenges and solutions of mental illnesses, such as Depression, are intuitively closely related to everyday difficulties, such as job, school, and even economic status. In contrast, Topic 4 appears to be quite distinct, focusing on darker subjects and a more inquisitive tone, marked by topics such as suicide, trauma, and looking for help, that don't necessarily align with the other discussions.

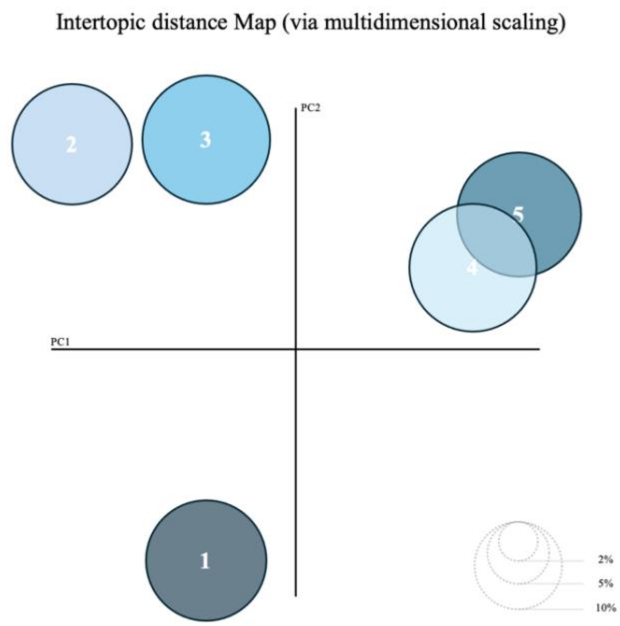


Exhibit 2: Intertopic distance map of the 5 topics identified through topic modelling.

To evaluate the quality and stability of the topics generated, we implemented an iterative process using different random seeds. A total of 100 models were created, each varying only by the random seed, to assess the consistency of the topics identified. Subsequently, a *doc2vec* model was employed to compare the topics produced by 50 of these models against the title and brief description of the topics from model 1. The findings are quite satisfactory: approximately 60% to 70% of the models showed a cosine similarity greater than 70% across all topics. This level of similarity is considered acceptable, particularly given that our analysis is primarily exploratory in nature.

5. Sentiment – Emotion analysis

The main objective of the analysis is to gain a deeper understanding into the effect of mental-health discussion on social media communities, in particular on reddit.

To carry out such analysis, we first looked at the topics identified in the exploratory stage and identified relevant related communities. From the topic of "Anger and Expletives", we focused on the subreddits "*r/Anger*" and "*r/angry*", from "Depression and Anxiety" on "*r/depression*", "*r/depression_help*," "*r/Anxiety*" and "*r/socialanxiety*", from "Stress in daily life" we focused on the subreddits "*r/Stress*" and "*r/hatemyjob*", lastly for the topic related to expressing emotions and worries to closed ones, we focused on the subreddit "*r/ToxicRelationships*". The data acquisition process was the same as that used for the general mental health-related subreddits, described in "*3. Data Acquisition*", collecting the top all posts.
Each subreddit consists of posts created by users on the main page, to which other users can respond with comments, hence starting a discussion. From this moment on, we will refer to comments as the reaction of other users' to posts, and replies/answers as the response to the comment, created by the original post creator. For each of these subreddits, the top posts for the year were collected as well as their respective comments and replies (*Exhibit* 3).

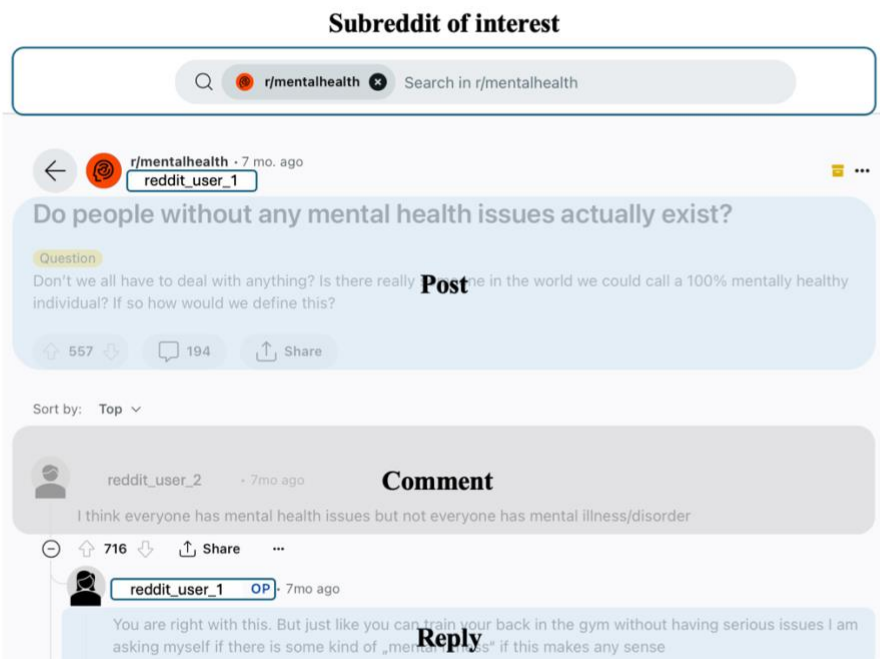


Exhibit 3: Structure of the Subreddits posts, comments, and replies.

The sentiment-emotion analysis was conducted in 2 main steps: the first step of the analysis is merely explorative, assessing the overall sentiment of the comprehensive dataset and within each subreddit, whereas the second step is related to the sentiment change in the reply after the comments.

5.1 The methodology

In both sections of the analysis the methodology employed is a pretrained BERT model, namely Twitter-RoBERTa-base for Sentiment analysis. This is a RoBERTa base model trained on about 58 million tweets and finetuned for sentiment analysis with the TweetEval benchmark proposed in "TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification" (Barbieri et Al. 2020). In this paper, it is proved that although Roberta base performs well on the majority of the tasks, Roberta RT, that is Roberta Base trained on the twitter corpus with the TweetEval benchmark, proves more effective (*Appendix* A).

Therefore, we decided to implement the code for RoBERTa Base trained on 58 million observation of twitter data, mainly due to 3 reasons:

1) It is a pretrained model on a large quantity of social media data, which is comparable to the reddit dataset of posts analyzed in our research;
2) It possesses easily interpretable and quantifiable labels for both emotion detection and sentiment analysis;
3) It is proved to be performing better than other BERT models on the task of sentiment analysis and emotion detection, which are our areas of interest.

For both of our tasks, we will therefore be implementing the model retrieved from the repository of the aforementioned paper, specifically the section about sentiment analysis and emotion detection[6].

We decided to employ emotion analysis, together with sentiment analysis, as strong emotions such as anger can often be mistaken as a negative sentiment, therefore we thought it was important to analyze deeper such emotions. The model evaluates each document according to the labels of the specific task. In sentiment analysis, the labels are three: "negative", "neutral", and "positive". In the case of emotion analysis, the labels are four: "joy", "optimism", "sadness", and "anger". The model then assigns for each document and for each label a score from 0 to 1. These scores are the model's confidence level: the higher the score, the more likely the model believes the text is expressing a particular label. Being standardized, they usually sum up to one and can be generally interpreted as probabilities. An example of the input and output format of the model is provided in the appendix (*Appendix* B).

The data preprocessing has been also inspired by the paper (removing the user and eventual website links), with the exception of also removing special characters (such as $, %, &, /, * etc.) with regular expressions, as they constituted an obstacle to model interpretation.

[6] The dataset and the complexity of the analysis carried out require a higher computing power than usually expected, we found it optimal to utilize a powerful hardware accelerator (in our case it is TPU v2 on Google Colab).

5.2 Exploratory analysis

In the first section, we first conducted an exploratory analysis of the overall sentiment and emotion of the subreddits of interest. After applying a simple preprocessing and the RoBERTa model for sentiment and emotion analysis, for each post we obtained a score for each label from 0 to 1.
Across all the datasets of all subreddits' posts, the overall sentiment is largely negative, as expected given the topics discussed (depression, anger, anxiety, toxic relationships, and anger). The negative sentiment is particularly high for posts about anger, of which more than 90% have the maximum score for negative sentiment, followed by posts about depression, anxiety, toxic relationships, and stress ranging from 70% to 80%. Positive and neutral sentiments are consistently the least common across these topics, indicating that individuals participating in these subreddits are generally expressing or discussing negative experiences or feelings. Regarding the theme of anxiety, however, over half of the posts have a negative sentiment, but almost 40% show the maximum score for positive sentiment, as shown in *Exhibit* 4. This indicates that the analyzed anxiety-related subreddits are not only used to seeking support and understanding, but also to share one's achievements, personal growth, strategies for overcoming challenges and lessons learned from their experiences.

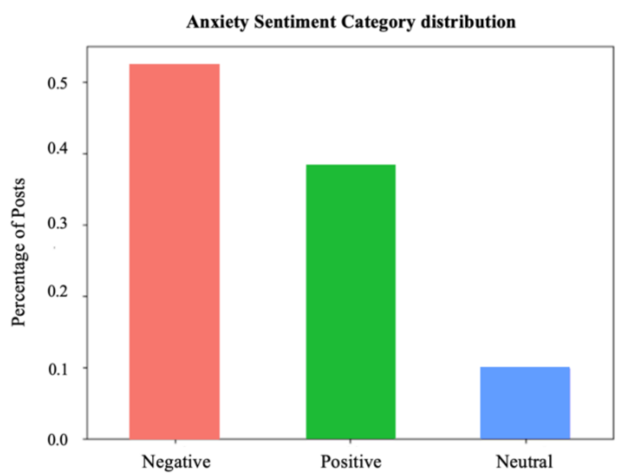


Exhibit 4: Sentiment category distribution in Anxiety posts.

All the other results are shown in the graphs reported in *Appendix D.*
In *Appendix D*, graphs related to the distribution and box plots of sentiment scores for each topic are also provided, confirming that all topics except anxiety have a strongly negative skewed distribution for negative scores. This indicates that a high number of posts have an almost sure negative sentiment. Additionally, there is a strongly positively skewed distribution for positive and neutral sentiment scores, which are thus nearly absent for the majority of posts.

To conduct a more in-depth investigation of the various topics, we carried out the emotion analysis. As shown in *Appendix D.7*, we obtained the following results:

- As expected, the predominant emotion in 80% of the posts about depression is sadness, followed by anger, while joy and optimism are almost entirely absent.
- Equally predictably, almost all of the posts in anger-related subreddits, around 90%, have anger as the predominant emotion.

- Consistent with the results obtained from sentiment analysis, 60% of the posts about anxiety predominantly express sadness, while 20% and 10% express positive emotions such as joy and optimism.
- In the subreddit dedicated to supporting people in toxic relationships, it is interesting to note that there is almost an equal percentage of posts, around 50%, where the predominant emotions are sadness and anger. Sadness may derive from disappointment and despair, while anger may arise from frustration in dealing in a toxic relationship, pushing individuals to seek support and solidarity in the online community.
- In subreddits related to emotional, social, and work-related stress, sadness and anger are the predominant emotions, covering almost 90% of the posts.

5.3 Comparative Analysis

In this section of the analysis the objective is to understand if and how the sentiment and emotions of the user are affected by other users' comments. To carry out this comparative analysis, we focused on a particular section of the dataset of all subreddits, where the posts present comment and answers. This is a much small subset of the overall posts, as it appears to be rather rare for users to respond to comments under their own posts, probably due to the sensitivity of the subject matter.

5.3.1 Frequency of labels in posts, comments, and replies

Across the whole dataset of subreddits' posts, comments, and replies, the sentiment and emotion analysis was conducted through RoBERTa pretrained model. Once again, we computed the score assigned to each label for each of the posts, comments, and replies. Then, we assigned to each post, comment, and reply, the label with the highest score, and compared them across predominant labels. The findings can be found in *Exhibit* 5.1 and 5.2.

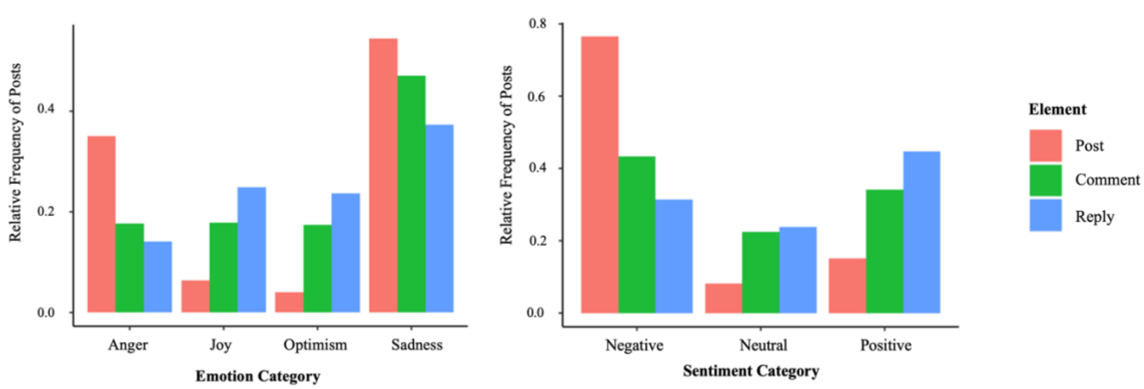


Exhibit 5.1 and 5.2: Frequency of sentiment and emotion labels.

Overall, posts (red colour in the graphs of *Exhibit* 5.1 and 5.2) tend to have a higher frequency of negative emotions and sentiment, almost 60% of the total number of posts showcases "sadness" as the dominant emotion, and 80% of the posts have a negative sentiment. Anger seems to also be the second most dominant emotion with almost 40% of the posts, whereas joy and optimism do not even reach 10% combined.

Comments (green colour in the graphs of Exhibit 5.1 and 5.2) follow a different distribution with respect to the posts, despite showcasing a predominantly large section of "sadness" (above 40%), the remaining 60% of comments appears to be distributed equally among "optimism", "joy" and "anger". The sentiment reflects in fact this trend, whereby only 40% of the comments have a negative sentiment, and an almost equal percentage appears to be "positive".
Surprisingly, a very similar distribution of the comments is followed by the replies (blue colour in the graphs of *Exhibit* 5.1 and 5.2), which seem to follow the same trend. "Sadness" remains the main emotion (below 40% of total replies) but followed by "joy" and "optimism", both well above 20%, a much improved percentage if compared to the 10% of the posts. This is also reflected in the prevailing sentiment: the "positive" portion of the replies even surpasses the "negative" one. This per se, already represents a significant result, as it proves the positive sentiment change of users from initial post to comments reaction.

5.3.2 Difference analysis filtering by type of Comment

Additionally, we decided to compute the average change in sentiment from posts to answers, comparing the average scores of each label. Overall, all types of comments seem to be increasing the positive and neutral scores, reducing the negative one from posts to replies. In *Exhibit* 6, we can see how the change in sentiment score (Answer – Post) changes across Comments' sentiment. It appears that all types of comments almost equally reduce the negative score but positive comments increase neutrality scores. However, neutral comments and negative comments appear to be stronger contributors to increasing the positivity score, which is a rather unexpected result.

Average Score difference (Answer - Post)				
Emotion	**All types of comments**	**Positive Comments**	**Negative Comments**	**Neutral Comments**
Positive	0.0851	0.0285	0.1116	**0.1201**
Negative	-0.2635	-0.2457	-0.2423	**-0.3313**
Neutral	0.1783	**0.2172**	0.1307	0.2112

Exhibit 6: Average score difference for Sentiment.

In order to explain such a result, we analyzed the change in emotions, as negative sentiment may be caused by different emotions: anger or sadness, both leading to very different interpretations.
From *Exhibit* 7 it is possible to compare the breakdown of the emotion difference analysis by type of comment. Combined, all comments seem to be contributing to the "joy" score, increasing the confidence of the model in detecting "joy" by 11 percentage points and "optimism" by 14.8 percentage points. The comments seem to be also reducing the model's confidence in detecting anger and sadness in replies by 4.1 and 2.2 percentage points respectively. This implies that overall, all types of comments seem to be inducing a positive sentiment change. However, comments characterized by a predominant "sadness" score seem to have the strongest effect on replies. Filtering for only "sadness" comments, the scores for both joy and optimism increase by the highest amount, namely 15 and 22 percentage points, whereas sadness decreases by the highest amount - 32 percentage points. These "sadness" comments, however, seem to not be effective in reducing anger, which is instead widely reduced in the case of optimism comments.

Average Score difference (Answer - Post)					
Emotion	**All types of comments**	**Joy comments**	**Anger Comments**	**Sadness comments**	**Optimism Comments**
Joy	0.1174	0.1463	0.1019	**0.1543**	0.0931
Anger	-0.0410	-0.0092	-0.0160	-0.0470	**-0.1335**
Sadness	-0.2249	-0.2151	-0.2464	**-0.3270**	-0.0775
Optimism	0.1485	0.0781	0.1604	**0.2197**	0.1179

Exhibit 7: Average Score difference for Emotion.

This also confirms our previous interpretation: both anger and sadness comments (comparable to negative sentiment) seem to be showcasing the highest reduction in sadness, which is in line with the strong reduction in negative sentiment in the negative comment subset of the data.

It is therefore possible that comments expressing empathy, sorrow and displeasure are followed by a more positive emotion change than joyful or optimistic comments. Optimistic comments and overall positive sentiment comments in fact seem to be performing quite poorly in increasing joy and optimism scores.

5.3.3 Difference analysis filtering by type of Post

Lastly, we analyzed if there was a difference across the different post types (*Exhibit* 8), keeping the filter for type of comment. The results appear different when specifying by post type. It is only in the subset of "sadness" posts that replies after "sadness" comments seem to be less likely to be marked as "sadness". However, still in that case, "joy" comments prove to be more effective in the task, reducing the "sadness" score by 57 percentage points.
In the rest of the cases, it is "joy" comments that witness an increase in "joy" score of the reply, and "optimism" comments that witness an increase in "optimism" score of the reply. Therefore, the previous results highlighting the increasing scores in "joy" and "optimism" due to "sadness" comments might have just been biased due to the high number of "sadness" posts in the dataset analyzed.

Average Score difference (Answer - Post)					
Posts	**Emotions**	**Joy comments**	**Anger Comments**	**Sadness comments**	**Optimism Comments**
Joy Posts	Joy	-0.1966	-0.3390	**-0.4292**	-0.3054
	Anger	0.0513	**0.1521**	0.1001	0.0509
	Sadness	0.0250	0.1488	**0.2092**	0.0357
	Optimism	0.1203	0.0381	0.1199	**0.2188**
Anger Posts	Joy	**0.3087**	0.1260	0.1334	0.1847
	Anger	-0.5514	-0.4402	**-0.5843**	-0.5779
	Sadness	0.0772	0.1800	**0.2889**	0.1907
	Optimism	0.1655	0.1341	0.1620	**0.2025**
Sadness Posts	Joy	**0.3452**	0.1147	0.1374	0.2351
	Anger	0.0544	**0.1498**	0.0565	0.0323
	Sadness	**-0.5704**	-0.4056	-0.3783	-0.5545
	Optimism	0.1708	0.1411	0.1845	**0.2871**
Optimism Posts	Joy	**0.3478**	0.0648	0.0528	0.2565
	Anger	0.0406	**0.1401**	0.0545	0.0386
	Sadness	**-0.0591**	0.1624	0.1216	-0.0116
	Optimism	**-0.3293**	-0.3674	-0.2289	-0.2835

Exhibit 8: Average score difference for type of Comment and Post (Answer score – Post score).

5.4 Results

From the analysis conducted, it is evident that overall the replies tend to have a more positive sentiment or to showcase more positive emotions than the original posts. We believe this may be correlated with the relief of expressing long held emotions or struggles with a community of users. In the initial phases of our analysis, filtering by comments only, we found that comments with a predominantly "sadness" label were usually followed by more positive replies. However, filtering also by post type, "joy" comments were found to be followed by replies with higher "joy" scores than "optimism" and "sadness" comments. In fact, it appears that for all types of posts, "joy" comments tend to be followed by replies with higher positive emotions' scores (both in optimism and joy). On the other hand, "sadness" comments appear to be followed by replies with higher "sadness" scores.

6. The Survey

Given the results in Section 5.4, we decided to assess whether the observed correlation among the comments, the emotion they express, and the change in the emotional tone from posts to replies could be considered causation. To do so, we conducted a survey comprising an interactive experiment.

We designed a questionnaire, in *Appendix E,* that is structured in 4 main sections, each with different goals carefully described in Section 6.1. The survey consisted of 20 close-ended questions and reached a total of 215 respondents. The results of the analysis are presented in Section 6.2.

6.1 The structure of the survey

General Information

The questionnaire initially gathers general information, such as sex, age, educational background, and current occupation. This data provides context, allowing analysis across various demographics. Then, respondents are asked if they have any family member who has or has experienced mental-health issues and if they have ever undergone psychotherapy. Such questions aim to construct a baseline of the respondent's mental-health landscape, which may influence how they perceive and are impacted by social media content.

Media Engagement

The second section of the questionnaire delves into the participant's social media usage. The questions aim to quantify their daily engagement in social media platforms and the main purposes of their usage. The emphasis then shifts to their interaction with mental-health-related content, asking the frequency of such engagements, their emotional connection with the content, and the perceived impact on their mental well-being. This section is pivotal as it directly correlates social media consumption patterns with the subjective mental-health states of individuals.

WHO-5 Well-being Index

In this part, general mental well-being of the respondent is assessed through the implementation of WHO-5 Well-being Index, a short self-reported measure of current mental well-being devised by the World Health Organization. Consisting of five questions, each offering six graded responses ranging from 0 (at no time) to 5 (all the time), the index yields a cumulative score that reflects the respondent's overall well-being in the last two weeks.
The questions/statements are:

1) *I have felt cheerful and in good spirits.*
2) *I have felt calm and relaxed.*
3) *I have felt active and vigorous.*
4) *I woke up feeling fresh and rested.*
5) *My daily life has been filled with thing that interest me.*

This index enables a quantitative assessment of mental well-being, allowing us to study how it changes across different social media consumption patterns, which has been observed in the prior section.

Interactive Experiment
The last section of the questionnaire involves an experiment in which participants are asked to assume the perspective of a social media content author, interpreting the emotional well-being level transmitted by a post.

"My life sucks. I'm 28 and I have no full time job and no passion. I have social anxiety and I really struggle having conversations with people. Moreover, I feel like I'm losing the only friend I have. I feel like an empty shell."

The respondents are then randomly divided into two groups, which receive two different treatments consisting in one of the two potential responses to this post.

1) A commiserative comment modelled to emulate the emotion "sadness": "*I can perfectly relate to what you're saying. I've been struggling with the same situation for years and to live with that is painful… you're not alone in this."*
2) An encouraging comment modelled to emulate the emotions "optimism" and "joy": "*I know you feel hopeless right now, but it will get better just by a little bit every day. Slow down and be kind to yourself. Stay strong and keep your head up!"*

Following exposure to one of these comments, respondents reassess their emotional well-being level. Our initial hypothesis is that the two distinct types of comments may cause a different impact on the readers.

6.2 Analysis

6.2.1 Data cleaning and pre-processing

Before employing any data analysis tool, it is crucial to undertake data cleaning and pre-processing to ensure the usability of the dataset. This process involved several key steps, listed below:

- The rows corresponding to respondents who did not complete the survey were removed.
- Some variables were recoded from character type to factor and numeric types as required.
- One-hot encoding was applied to variables from multiple-answer questions; single-answer questions are automatically encoded by the modeling software.
- The variables with overly long labels, originally based on survey choices, were renamed using more concise identifiers.
- For the WHO-5 Well-being Index, we assigned a numeric value (0 to 5) to each response and computed a total well-being score, which was stored as a new variable.
- For the variable containing the reassessed emotional well-being level of the respondents, we had the responses of the units assigned to Treatment 1 in one column and the responses of the units assigned to Treatment 2 in another column, both having as many NA values as many respondents were assigned to the other treatment. The two columns were combined into one single column (named "After"), while retaining the variable informing about the specific treatment received by each respondent.
- A new variable was created to represent the change (delta) in emotional well-being by computing the difference between the variables "After" (reading the comment) and "Before" (reading the comment).

After conducting an initial exploration of the data and computing summary statistics, we identified six potential outliers. These included:

1) Five respondents who, when asked to imagine themselves as someone in obvious distress, rated their well-being at the maximum possible value (10) on a scale of 1 to 10.
2) One respondent who experienced an eight-point drop in his well-being rating (out of 10) following a supportive comment.

To preserve the integrity of our dataset, we decided to exclude these observations from further analysis.

6.2.2 Exploratory Data Analysis

Sample Demographics

In the initial step of our analysis, we conducted a deep study of the sample demographics, and the key characteristics are reported below.
The majority of the sample consists of female participants, making up slightly more than half of all respondents. Approximately 78% of the participants fall within the 18-25 age range, indicating a predominantly young demographic. About 73% of respondents are students, while approximately 19% are employed in full-time roles.

Mental-health History and Psychotherapy Treatment

We then shifted our attention to factors concerning familial history of mental-health issues and experiences with psychotherapy.
Around 45% of participants disclosed having a family member who has or has experienced a mental-health issue, while 28.5% have undergone psychotherapy. The mosaic plot shown in *Exhibit*

9 revealed a clear correlation: individuals with a family history of mental-health issues tend to seek therapeutic assistance more frequently; conversely, those without such familial history show a lower inclination towards therapy.

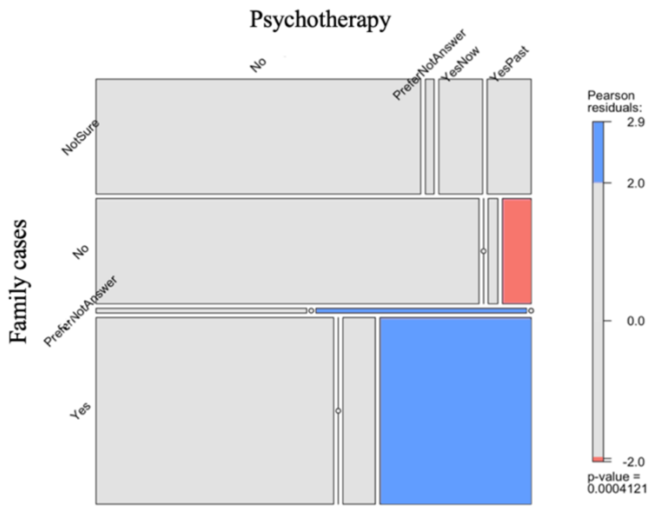


Exhibit 9: Mosaic plot showing the relation between family history of mental-health issues and psychotherapy treatments.

Patterns of Social Media Usage

The next phase of survey analysis aimed to find diverse patterns of social media usage. As we can see in *Exhibit* 10, just a small fraction f the sample, approximately 6%, spends 1 hour or less per day on social media platforms. Remarkably, around 61% of respondents allocate more than 3 hours daily to these platforms, with over 20% using them for more than 4 hours each day. Instagram emerged as the preferred choice for the majority of the sample, with approximately 66% favoring it, while 26% expressed a preference for TikTok as their primary social media platform. The preference for these two platforms suggests that visual content is particularly engaging.

We deduce that social media are a prominent part of the daily life for the majority of the survey participants. Given the amount of time is spent on these platforms, there’s a potential for social media to have a considerable impact on individuals’ daily routines and potentially their mental well-being.

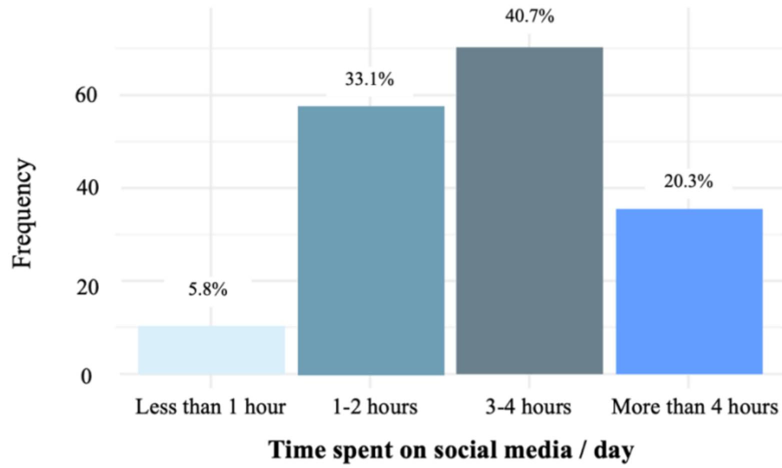
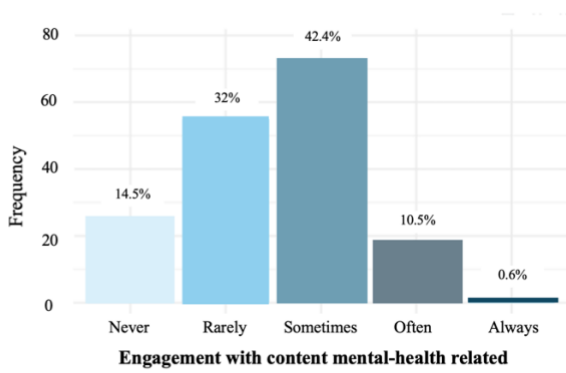
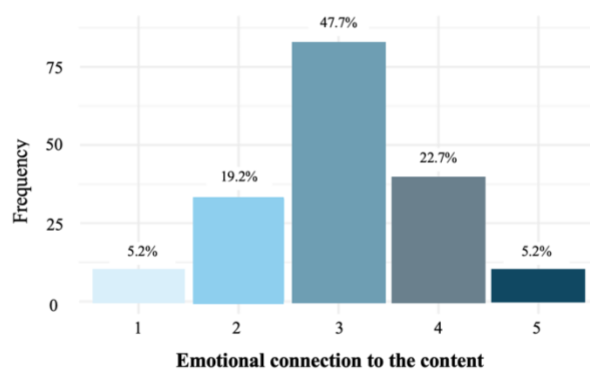


Exhibit 10: Barplot showing absolute and relative frequency for the variable representing the time spent on social media per day.

Mental-health related content

As shown in *Exhibit* 11.1, approximately 42% of participants occasionally engage with content related to mental-health, suggesting that there is a notable interest in such content among social media users. Emotional involvement with mental-health content (*Exhibit* 11.2) exhibits a symmetrical distribution centered around the midpoint value of 3: 48% of respondents rate their emotional connection at the middle level, while 28% report high emotional engagement, scoring 4 or 5 out of 5 on the scale. The impact of mental-health content on the respondents' well-being, shown in *Exhibit* 11.3, displays a left-skewed distribution: roughly 54% of survey participants view it positively. The left-skewed distribution suggests that the impact of mental-health content on respondents' well-being is generally positive, with more than half of the participants perceiving a beneficial effect.





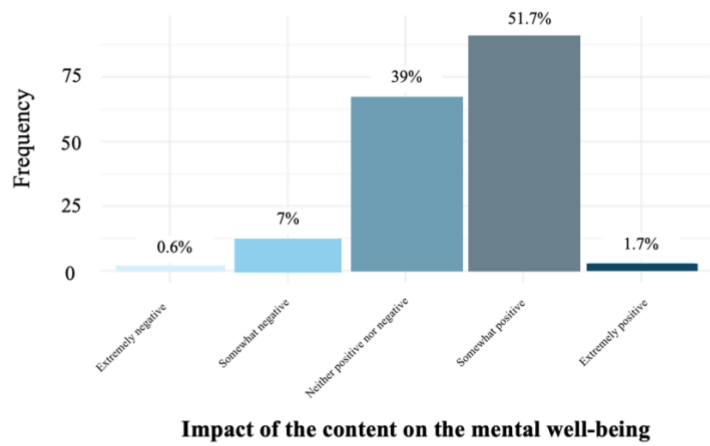


Exhibit 11.1, 11.2 and 11.3: Barplots showing absolute and relative frequency for the variable representing the engagement of the users with content mental-health related (11.1), the emotional connection of the users with this kind of content (11.2) and the perceived impact of the content on the user's mental well-being (11.3).

The WHO-5 Well-being Index

As last step of the exploratory analysis we analyzed the WHO-5 Well-being Index score and its relation to other variables. The distribution, shown in *Exhibit* 11, has its mean value in 12.6 and the median in 13. The slight discrepancy between the mean and median indicates a minor negative skew in the distribution of scores, suggesting a good mental well-being among most participants.

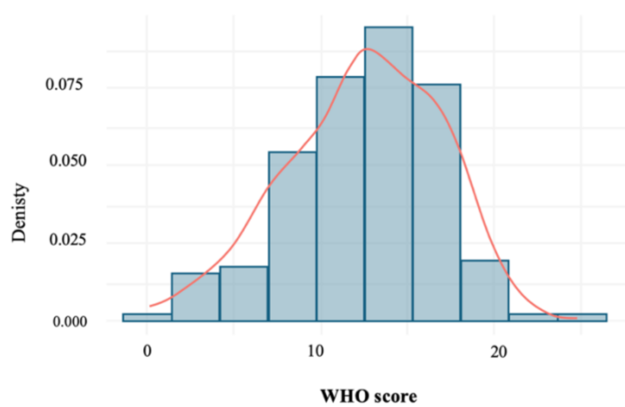


Exhibit 12: Histogram and density of the distribution of the WHO-5 Index.

Additionally, a trend emerged linking extensive social media usage with lower WHO-5 scores (*Exhibit* 13.1), suggesting that heavy use of social media could be associated with poorer mental well-being. Participants with a strong emotional connection to social media content tend to exhibit lower WHO-5 scores (*Exhibit* 13.2), suggesting that those who are more emotionally affected by social media may experience more negative emotional states or stress, which in turn could affect their overall well-being.

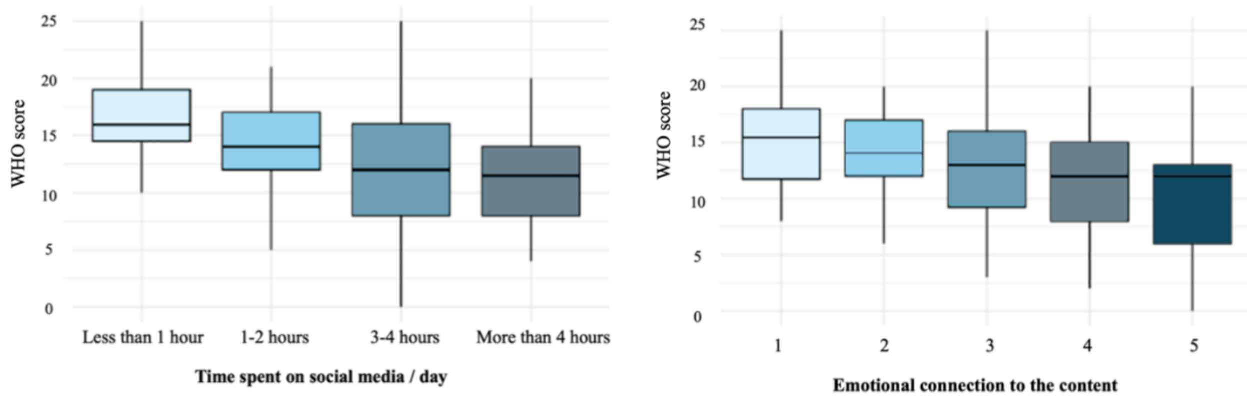


Exhibit 13.1 and 13.2: Boxplots of the distribution of the WHO-5 Index conditioned on the time spent on social media per day (13.1) and the emotional connection of the user with the mental-health related content seen on social media (13.2).

6.2.3 Experiment: the effects of the comments of the users

The third and last part of the analysis was dedicated to the effect of the two different kinds of comments on the state of emotional well-being level assessed by the respondent before and after reading it.

A balance table verified that the experimental groups were comparable across all measured covariates, establishing that any differences in outcomes are likely due to the treatments themselves rather than pre-existing differences between the groups.

Despite our initial hypothesis that the two distinct types of comments (in Section 6.1 - *Experiment*) might have different impacts on readers, the statistical analysis conducted did not provide evidence to support this. Specifically, it appears that the nature of the comments does not lead to a statistically different effect on the readers' responses. Note that the effect was computed as the difference between the variables assessing the emotional well-being level of the respondent "After" (reading the comment) and "Before" (reading the comment).

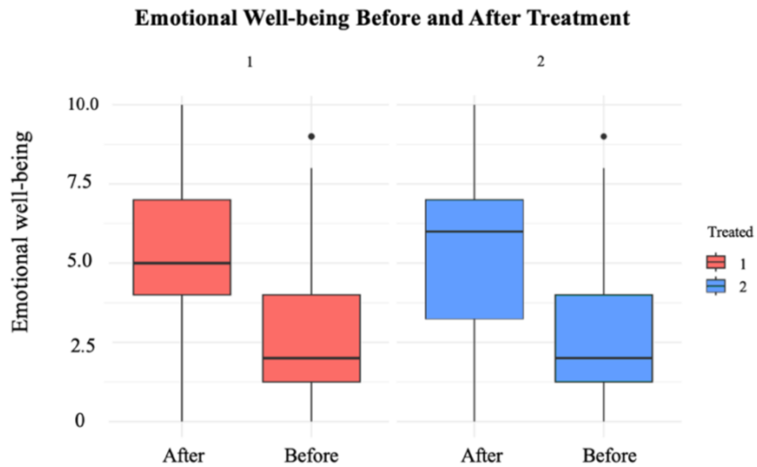


Exhibit 14: Boxplots of the distribution of the emotional well-being level of the respondent conditioned on the timepoint (before and after the treatment) and on the treatment (the comment shown).

However, when examining the well-being scores of respondents before and after reading the comments as a whole, without distinguishing between comment types, there is a notable change. The analysis shows a statistically significant difference in the scores, with an average shift of 2.46 points. The presence of this statistically significant difference underscores the impact that engaging with comments can have on an individual's state of mind, highlighting the potential influence of social media interactions on well-being.

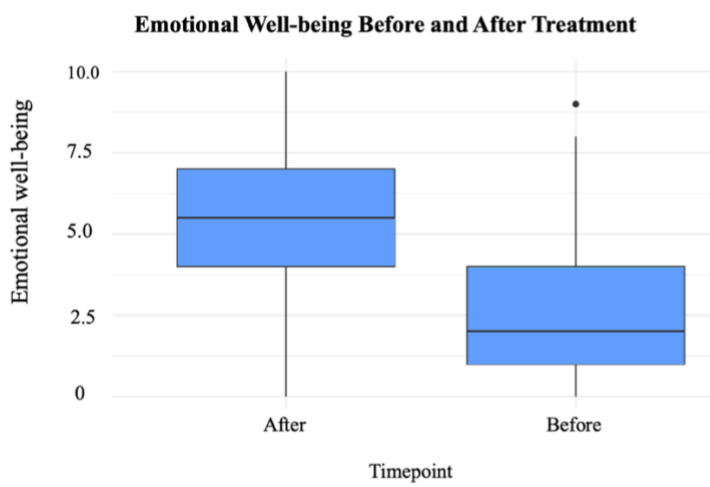


Exhibit 15: Boxplot of the distribution of the distribution of the emotional well-being level of the respondent conditioned on the timepoint (before and after the treatment).

The last result can be causally interpreted. We can consider the experiment setting under a Before-After Experiment point of view, in which the change in the emotional well-being level assessed by the respondent can be only due to the comment read by them. However, we need to take into account that this is a lab experiment, in which there could be external validity concerns and the Hawthorne effect[7] may play a role.

7. Conclusions

The objective of our study was to assess the extent to which social media communities dedicated to the discussion of mental-health topics can impact an individual's mental well-being.

From our exploratory empirical analysis, it appears that the majority of the posts related to mental health communities tend be dominated by a negative sentiment, confirming our hypothesis of individuals turning to online communities for help and understanding. This may be due to the taboo that usually accompanies such sensitive topics, but also due to issues of care accessibility, either geographical or economical. A positive correlation between the users' interaction (on dedicated communities) and improved emotional wellbeing was also found, as the sentiment improves after said exchange. The experiment confirmed our findings: receiving some sort of feedback, regardless of it being complacent or encouraging, has the effect of improving the morale of the user posting, and those observing such exchange.

It is therefore clear that social media platforms and communities have now become a reference point for seeking help and comfort, due to the mutual support created by users' shared experiences. This study demonstrates that, in conjunction (and not in substitution) to proper mental health care, social media communities can have a positive impact on users' mental well-being.

In essence, this investigation not only proves the potentially positive role of digital communities in mental-health promotion, but also contributes to informing policymakers and health professionals about the potential of social media as a tool for public health intervention, especially in contexts where traditional mental-health services are either unavailable or prohibitively expensive.

[7] The Hawtorne effect refers to the fact that the respondents know they are under evaluation, and this condition forces them to change their behaviour.

8. Bibliography

1. It states that the more time humans spend on social media, the less time they spend in the real world.
2. Naslund, J.A., Bondre, A., Torous, J. et al. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. J. technol. behav. sci. 5, 245–257 (2020). https://doi.org/10.1007/s41347-020-00134-x
3. Palis, H., Marchand, K., & Oviedo-Joekes, E. (2020). The relationship between sense of community belonging and self-rated mental health among Canadians with mental or substance use disorders. Journal of Mental Health, 29(2), 168–175. https://doi.org/10.1080/09638237.2018.1437602
4. https://www.who.int/news-room/fact-sheets/detail/depression
5. Patel V. Mental health in low- and middle-income countries. Br Med Bull. 2007;81-82:81-96. doi: 10.1093/bmb/ldm010. Epub 2007 Apr 30. PMID: 17470476.
6. Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). TWEETEVAL: Unified benchmark and comparative evaluation for tweet classification. Snap Inc., Santa Monica, CA; School of Computer Science and Informatics, Cardiff University, United Kingdom. https://aclanthology.org/2020.findings-emnlp.148.pdf and repository link: https://github.com/cardiffnlp/tweeteval

9. Appendix

Appendix A: Performance of RoBERTa RT Model compared to RoBERTa Base and Twitter

		Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	ALL
Val	SVM	25.0	63.8	73.1	63.4	72.7	68.4	67.9	62.0
	FastText	23.2	62.9	71.7	62.7	70.0	62.2	67.3	60.0
	BLSTM	19.4	62.6	72.1	60.6	72.1	61.9	63.4	58.9
	RoB-Bs	24.7±0.3 (24.3)	73.1±1.7 (74.9)	76.5±0.3 (76.6)	73.7±0.6 (73.7)	77.1±0.6 (77.6)	71.4±1.9 (72.7)	71.4±1.9 (73.9)	67.7
	RoB-RT	24.4±1.5 (**26.2**)	75.4±1.5 (**77.0**)	77.8±1.1 (**79.6**)	74.7±1.5 (**75.6**)	77.2±0.6 (**77.7**)	73.0±1.2 (**74.2**)	72.9±1.0 (**75.2**)	**69.4**
	RoB-Tw	23.4±1.1 (24.6)	67.6±0.9 (68.6)	74.3±2.0 (76.6)	70.0±0.3 (70.7)	76.1±0.6 (76.2)	70.5±1.0 (69.4)	68.3±2.4 (71.4)	65.4
Test	SVM	29.3	64.7	36.7	61.7	52.3	62.9	67.3	53.5
	FastText	25.8	65.2	50.6	63.1	73.4	62.9	65.4	58.1
	BLSTM	24.7	66.0	52.6	62.8	71.7	58.3	59.4	56.5
	RoB-Bs	30.9±0.2 (30.8)	76.1±0.5 (76.6)	46.6±2.5 (44.9)	59.7±5.0 (55.2)	79.5±0.7 (78.7)	71.3±1.1 (72.0)	68±0.8 (70.9)	61.3
	RoB-RT	31.4±0.4 (**31.6**)	78.5±1.2 (**79.8**)	52.3±0.2 (**55.5**)	61.7±0.6 (62.5)	80.5±1.4 (**81.6**)	72.6±0.4 (**72.9**)	69.3±1.1 (**72.6**)	**65.2**
	RoB-Tw	29.3±0.4 (29.5)	72.0±0.9 (71.7)	46.9±2.9 (45.1)	65.4±3.1 (**65.1**)	77.1±1.3 (78.6)	69.1±1.2 (69.3)	66.7±1.0 (67.9)	61.0
	Best	36.0*	-	65.1	70.5	82.9	68.5	71.0	-
Metric		M-F1	M-F1	M-F1	$F^{(i)}$	M-F1	M-Rec	AVG ($F^{(a)}$, $F^{(f)}$)	TE

Source: “TWEETEVAL: Unified benchmark and comparative evaluation for tweet classification” (Barbieri et. Al. 2020)

Appendix B: Example of the RoBERTa model input and output structure

Input: “Good Night!”

Output:

1) positive 0.8466

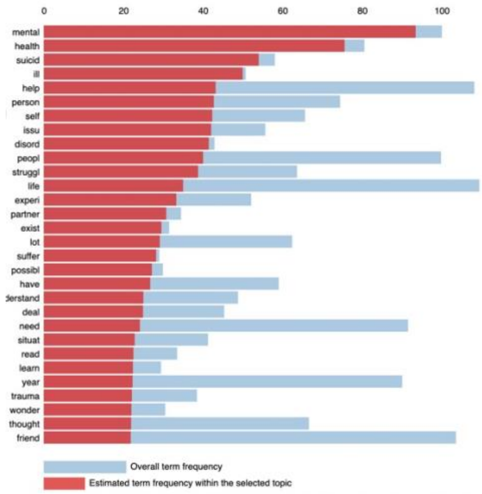
2) neutral 0.1458

3) negative 0.0076

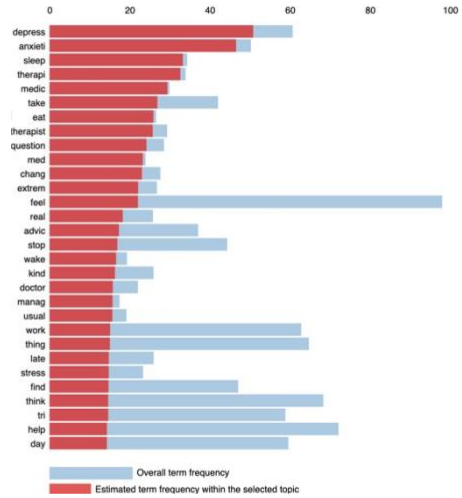
Source: “TWEETEVAL: Unified benchmark and comparative evaluation for tweet classification” (Barbieri et. Al. 2020) repository: https://github.com/cardiffnlp/tweeteval

Appendix C: Top 30 most relevant terms for each topic

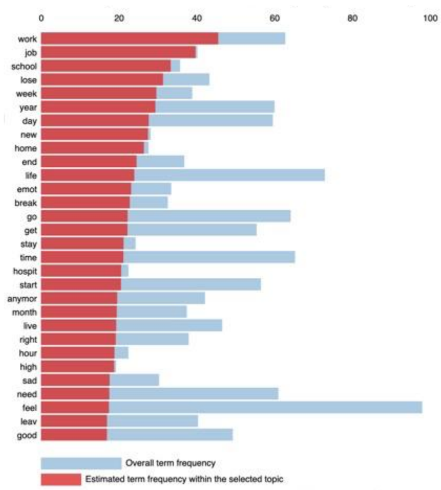
Topic 1



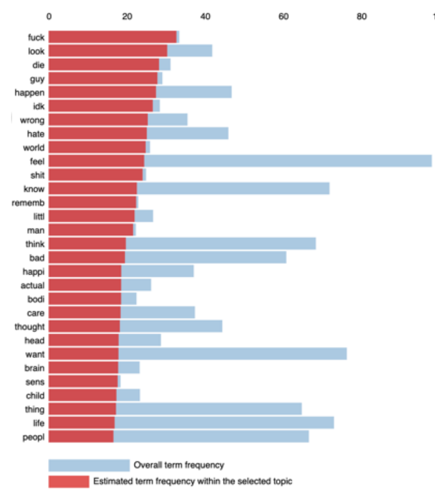
Topic 2



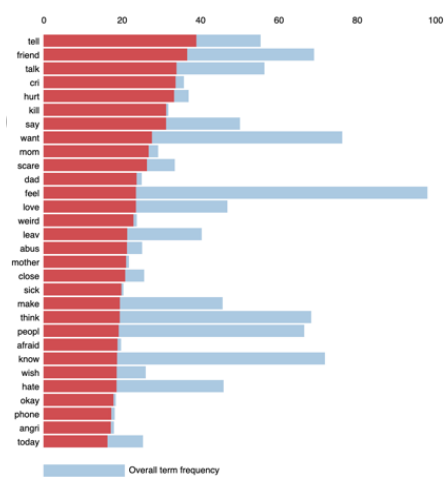
Topic 3



Topic 4



Topic 5



Appendix C.1: Top 3 most correlated posts for each topic

Topic 1:

1) *"My husband (m37) self-diagnosed himself with ADHD and depression. I (f36) do believe his behaviour is of someone with mental health challenges. I've asked him to seek proper diagnosis as that is the only way to deal should it be the case. I no longer know how to calm him or communicate with him without it being made into conflict as he's diagnosed me with narcissism when I call him out for issues or anything. His family doesn't help the situation as they seek to make me a bad person and enable some of the choices he makes. Help what is the best way forward?"*
2) *"I have had a very negative experience with mental health professionals diagnosing me with an illness I do not have and very aggressive treatment from mental health professionals. Wondering if there are others out there having the same issue?"*
3) *"I feel embarrassed about how many mental illness I have. I have been diagnosed with several mental illness. any time I talk about it I feel like a disgusting attention ***** or a pick me. especially because a few of those diagnosis are pretty severe mental health issues."*

Topic 2:

1) *"Just done. . . I have tried everything and nothing helps"… "I just slowly and steadily feel worse and worse everyday. Nothing motivates me. Nothing interests or entertains me. I have no goals, no dreams, no desires nor aspirations. I don't get enjoyment out of anything."*
2) *"I have no idea what more I can do. Anxiety?"…"It's a pervasive worry that's not about catastrophic events but more personal concerns—like questioning whether my body and movements appear normal."… "This worry consumes me; I'm constantly battling to maintain composure and conceal my struggles from my colleagues and managers."*
3) *"My psych seems less interested in me now and almost frustrated when we meet. I'm worried she'll drop me as a patient"… "Been prescribed various medications but not sure I ever knew my diagnosis? I guess just anxiety and depression?"*

Topic 3:

1) *"Feeling like I want my mum I (27F) live 3 and a half hours away from my parents and all family. I moved up for a job that I have been dismissed from. Been fine for 3 months even though I'm unemployed."… "Is it because I'm alone again and missing that connection I had with her?"*
2) *"My life is falling apart So i 32M left home to chase an opportunity because my wife 31F and I were struggling so bad financially and i also had a mental break. She doesnt work because she's disabled, so now almost a year later I'm 1500 miles away and i found out she's cheating on me, my house is trashed, my bank account is basically empty."*
3) *"At the age of 35, I am in the best shape of my life physically and the worst shape mentally."… "The place I was working for two years closed down and I have no desire or energy to find something else at the moment. That was also the longest consecutive amount of time I have been employed since graduating high school."*

Topic 4:

1) *"Help me I hate everyone It's not just hate, its disgust. I cannot stand this world, the people in it. Oh how I want to just burn this world and the people in it to ashes. I ******* hate to see people happy and smiling"… "Things just have to go worse and worse and worse. I know there are many people who also it way worse than me"… "what do I do with my rage and hatred?"… "Because it's so*

******** explosive, I feel my head would burst I might just do something to someone. Most probably the closed ones"*

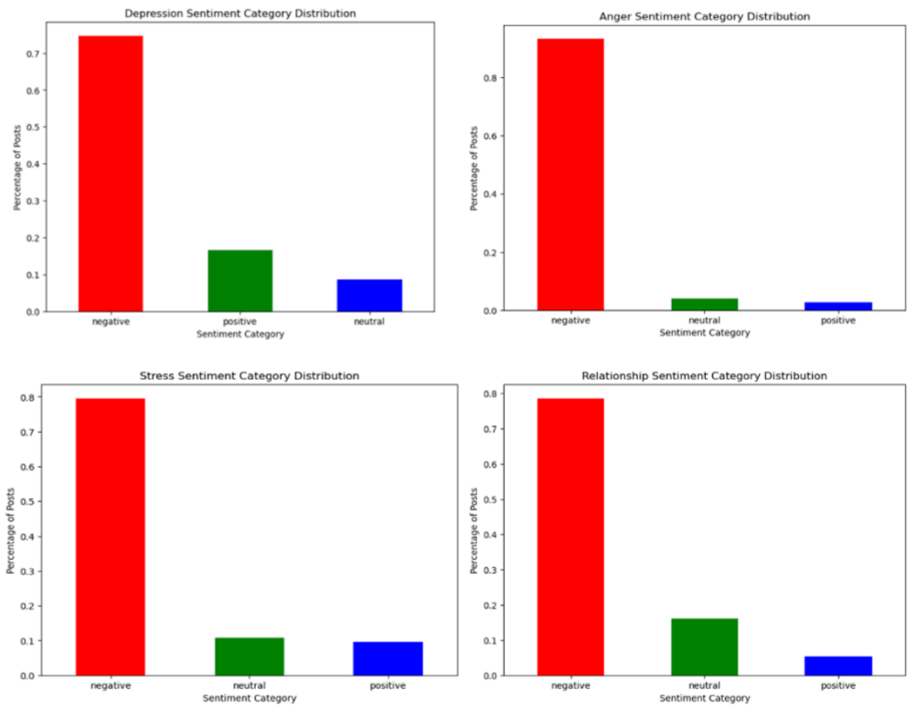
2) *"Idk what is wrong with me i just want to d** I cant bear living any longer" … "i cant do it anymore".*
3) *"lol im so ******* cringe **** man i feel so wrong. like im nothing and i dont matter everything i do is s***** and i look s***** and i hate everything about myself."*

Topic 5:

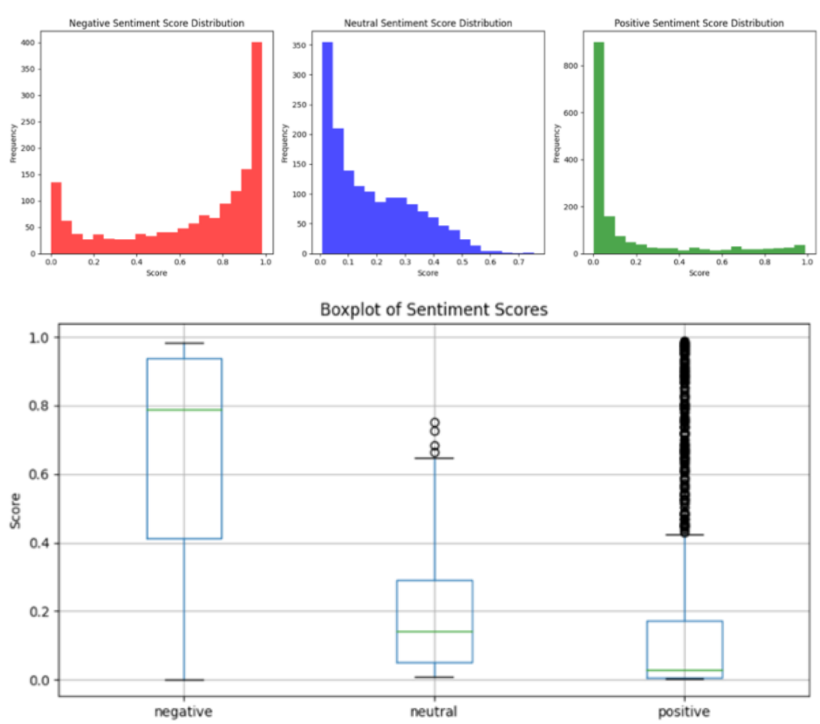
1) *"I'm so tired I'm tired of feeling like I'm invisible to everyone. I decided today to see if anyone would talk to me if I didn't start the conversation, and guess what. No one did."*
2) *"No one wished me happy birthday today I feel like I'm bad and unimportant I thought I was special to some of my friends I don't have any friends. only my close family wished me"*
3) *"toxic friend.." … "She never like gives me advice, for example I told her that I was thinking suicidal thoughts and she just said 'stop that' and that was it, she always goes to me for advice" … "I've had to suffer with this girl bringing down my mental health for four years now and I don't even like her"*

Appendix D: Emotion and Sentiment exploratory analysis of the posts

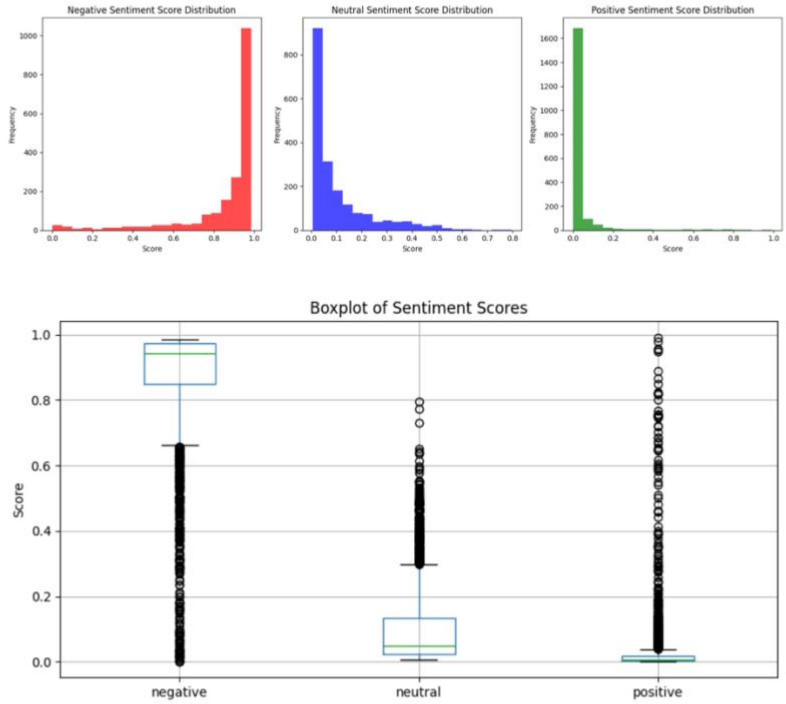
Appendix D.1: Boxplot of Sentiment Categories for each topic



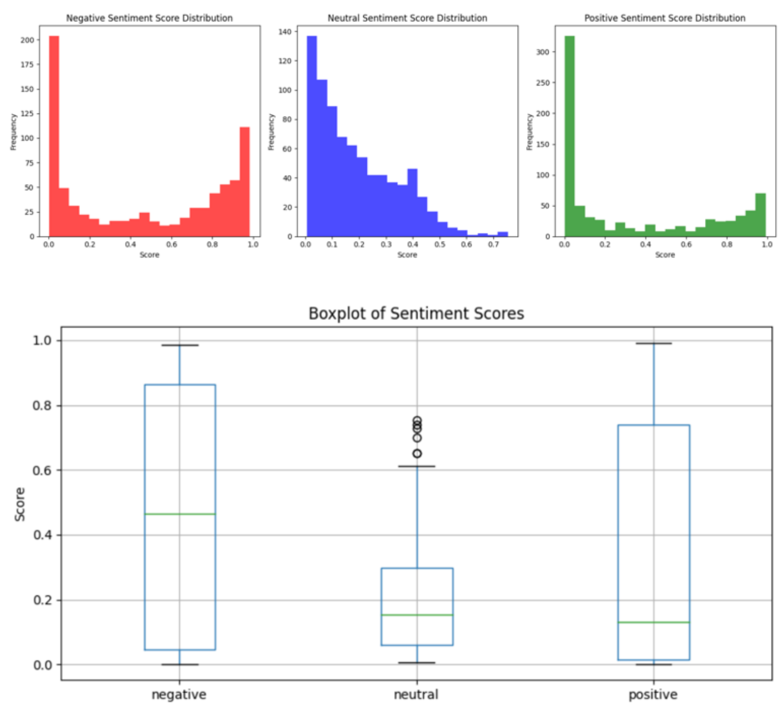
Appendix D.2: Depression Sentiment Score Distribution



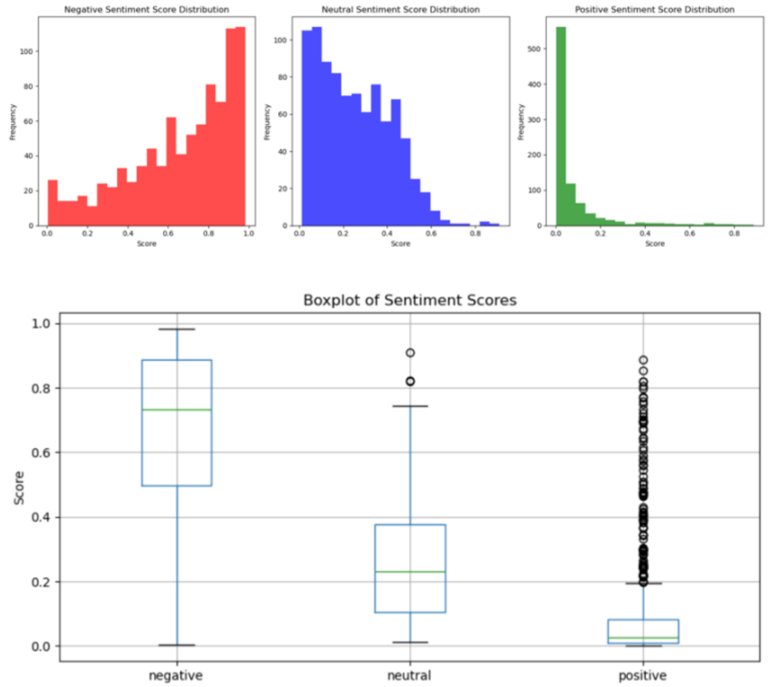
Appendix D.3: Anger Sentiment Score Distribution



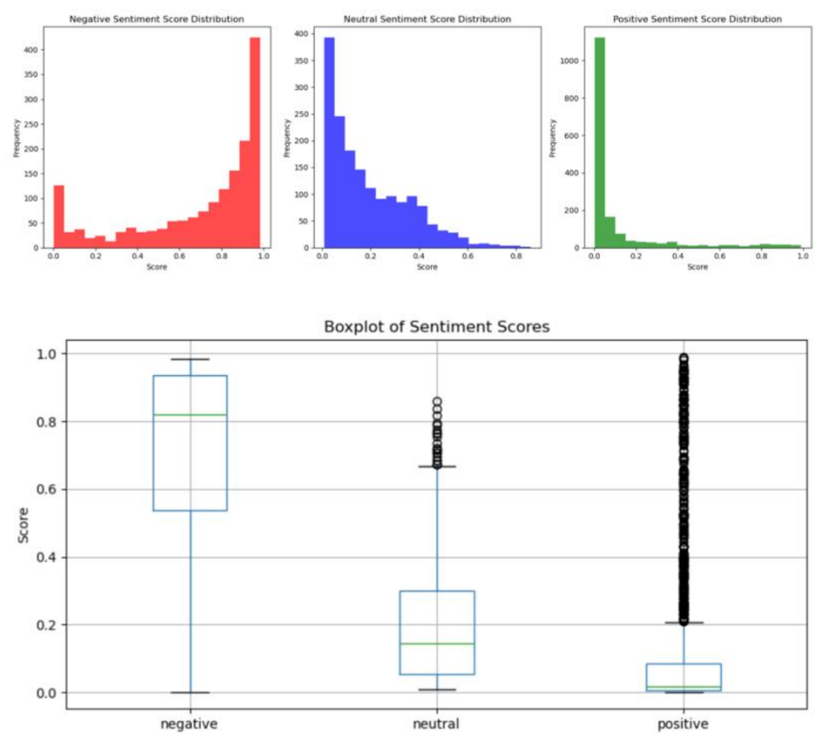
Appendix D.4: Anxiety Sentiment Score Distribution



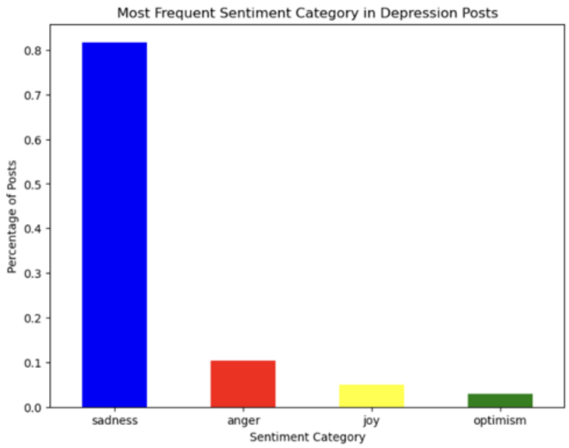
Appendix D.5: Relationship Sentiment Score Distribution

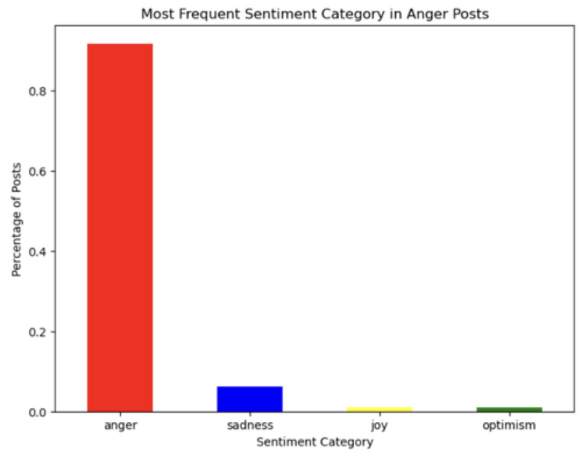


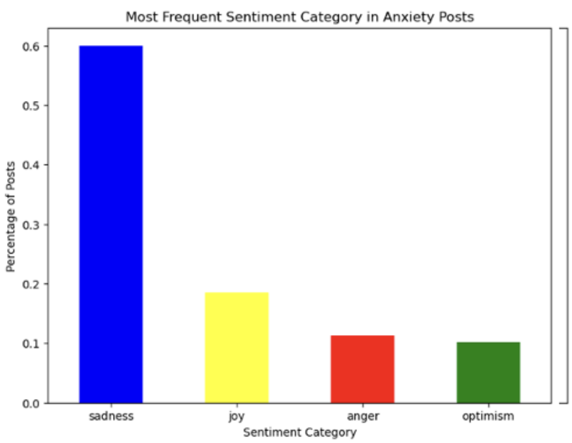
Appendix D.6: Stress Sentiment Score Distribution

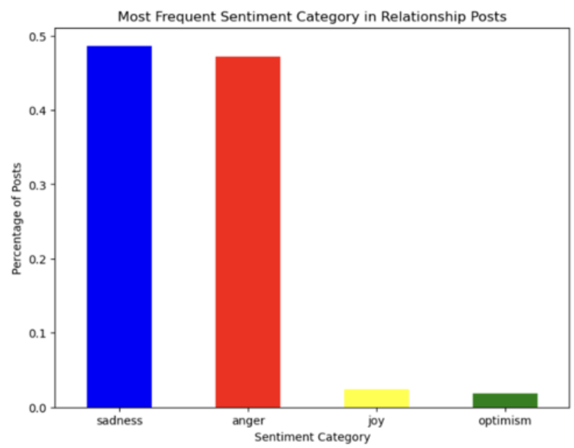


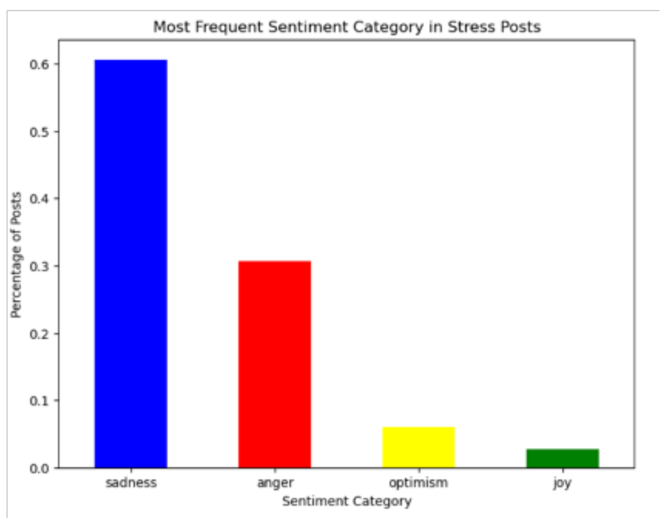
Appendix D.7: Boxplots of Emotion Categories for each topic











Appendix E: The survey

Do you have any family members who have or have had mental health issues such as anxiety, depression, etc.?

- Yes
- No
- I'm not sure
- Prefer not to answer

Have you ever received psychotherapy treatment?

- Yes, I have received psychotherapy treatment in the past
- Yes, I am currently undergoing psychotherapy treatment
- No, I have never received psychotherapy treatment
- Prefer not to answer

How many hours per day do you spend on social media?

- Less than 1 hour
- 1-2 hours
- 3-4 hours
- More than 4 hours

Which social media platform do you use the most?

- Facebook
- Instagram
- TikTok
- LinkedIn
- X (Twitter)
- Reddit
- Other

For what purposes do you use social media? Multiple answers allowed.

- ☐ Connecting with friends and family
- ☐ Networking
- ☐ News and information
- ☐ Entertainment
- ☐ Support and advice
- ☐ Education and learning
- ☐ Hobbies and interests
- ☐ Creative expression
- ☐ Other

How often do you engage with posts, videos, comments, and pages on social media regarding mental health (i.e., anxiety, depression, self-confidence issues, burnout etc.) ?

- ○ Never
- ○ Rarely
- ○ Sometimes
- ○ Often
- ○ Always

What kind of activities do you engage in with these contents? Multiple answers allowed.

- ☐ Reading and watching
- ☐ Commenting
- ☐ Sharing content
- ☐ Seeking advice
- ☐ Offering advice
- ☐ Other

On a scale from 1 (not at all) to 5 (extremely), how emotionally connected do you feel to these contents?

1	2	3	4	5
○	○	○	○	○

What is the effect of mental health-based content on your mental well-being?

- ○ Extremely negative
- ○ Somewhat negative
- ○ Neither positive nor negative
- ○ Somewhat positive
- ○ Extremely positive

English

Please respond to each item by marking one box per row, regarding how you felt in the last two weeks.

	All of the time	Most of the time	More than half the time	Less than half the time	Some of the time	At no time
I have felt cheerful and in good spirits	○	○	○	○	○	○
I have felt calm and relaxed	○	○	○	○	○	○
I have felt active and vigorous	○	○	○	○	○	○
I woke up feeling fresh and rested	○	○	○	○	○	○
My daily life has been filled with things that interest me	○	○	○	○	○	○

The Treatment:

Imagine being the author of the following post written on a mental health-based discussion in an online community. Please read it carefully.

"My life sucks. I'm 28 and I have no full time job and no passion. I have social anxiety and I really struggle having conversations with people. Moreover, I feel like I'm losing the only friend I have. I feel like an empty shell."

As the author of the post, please rate on a scale from 0 (severe pathological level) to 10 (normal emotional state) where your level of emotional wellbeing would be at.

0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

English

Picture yourself reading the following comment to your post:

"I know you feel hopeless right now, but it will get better just by a little bit every day. Slow down and be kind to yourself. Stay strong and keep your head up!"

After reading the comment, think about the impact that it had on your emotional well-being. Please rate on a scale from 0 (severe pathological level) to 10 (normal emotional state) where your level of emotional wellbeing would be at.

0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○