Depression subreddits text analysis

Jenny Gong

Summary

The surge in posting activity coincided with the onset of the COVID-19 pandemic, a period marked by increased self-doubt among individuals with depression. Contributing factors to this sentiment may include occupational stress, deteriorating interpersonal relationships, and challenges in romantic partnerships.

Research objective

Depression is a widespread negative emotional condition, and it can result in self-destructive actions and occasionally suicide (Tejaswini, 2022). Initial data indicate that the overall prevalence of depression has been worsened by the COVID-19 pandemic. In 2020, nearly 10% of Americans and almost 20% of adolescents and young adults experienced depression within the past year. Young adults in the age group of 18 to 25 years exhibited a higher prevalence of depression at a rate of 17.2% (Goodwin, 2022).

Considering the persistent high prevalence of depression among young adults, who are the predominant users of social media, this study seeks to analyze social media posts. Through the training of a social media corpus, the research aims to enable early depression screening and intervention, thereby preventing the condition from advancing to a severe stage.

This research aims to explore the topics that individuals impacted by depression discuss. It employs natural language processing techniques to uncover prominent themes in their conversations. It seeks to unveil how they articulate their emotions and potentially discern variations in the expression of depression, including subtypes and important tokens. Following this exploration, the objective is to aggregate these findings to construct a model serving as an early intervention tool.

Data sources

The dataset was sourced from Reddit, encompassing a total of 59,242 posts spanning the timeframe from November 2018 to April 2020, missing the data from December 2019. The data resources employed in this study were derived from the comprehensive mental health research, utilizing natural language processing techniques (Low, 2020). Low's research encompassed a comprehensive analysis of various mental health-related subjects, such as anxiety, ADHD, autism, and more. However, for the purposes of this study, our focus was specifically on the subreddit dedicated to discussions about depression.

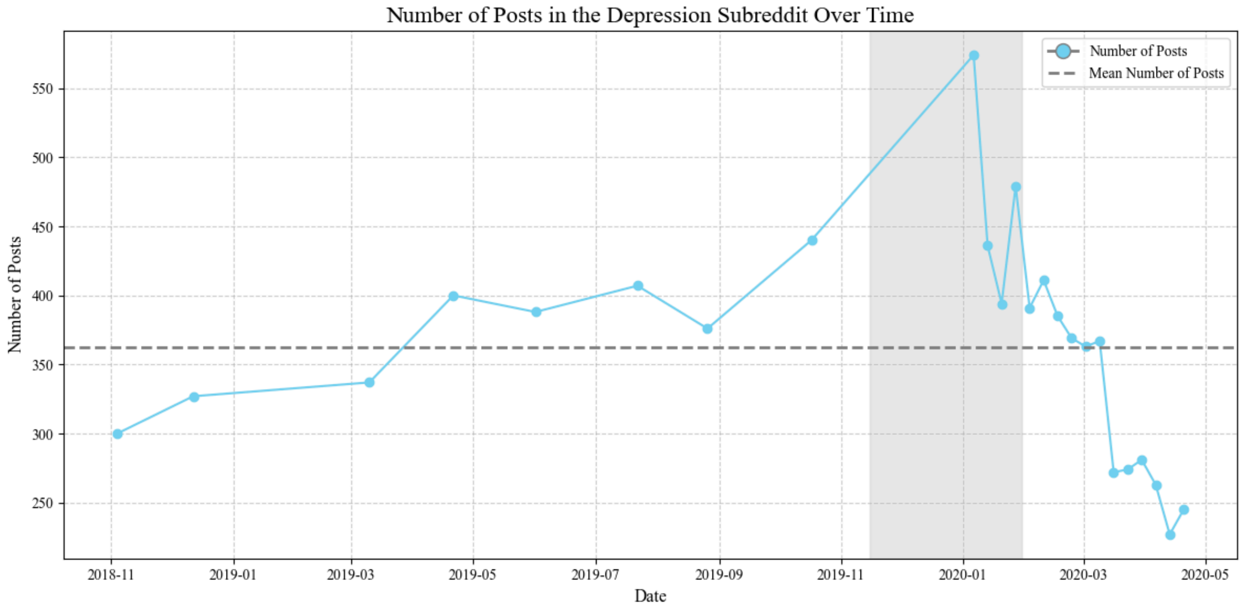

Figure.1 The number of posts overview

The volume of posts exhibited a high frequency preceding February, likely attributed to the onset of the pandemic in January. Nevertheless, the overall trajectory indicates a declining trend. Post-mid-March, the daily count consistently remained below the threshold of 300. This reduction in posting activity may reflect the evolving dynamics of the pandemic and may result from multiple factors.

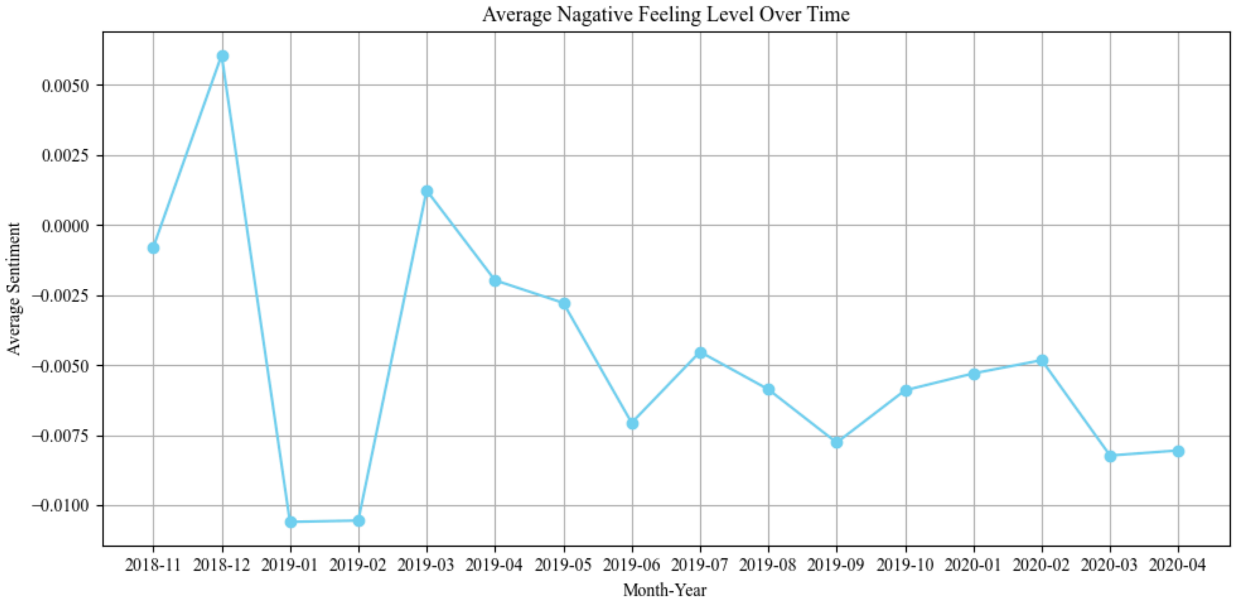

Figure.2 Average negative feeling level over time

The sentiment analysis graph shows sentiment trends from November 2018 to April 2020. The y-axis represents average sentiment scores, with values closer to zero indicating less negative sentiment and values further below zero indicating increased negativity. Initially, sentiment is volatile, with a sharp decrease in negativity in December 2018 followed by a spike in February 2019.

Over time, sentiment moderates, with peaks and troughs becoming less pronounced, suggesting stabilization in negative sentiment expression. This may reflect community adaptation or changing reactions to external events. The latter half shows a gradual decline in negative sentiment, especially from August 2019 to February 2020, with a notable uptick in April 2020. Context, such as major events or user demographics, is needed to explain this fluctuation. Nevertheless, the data offers valuable insights into the collective emotional journey of the subreddit's users during this period.

Limitation of the dataset

In the context discussed earlier, it's important to note that the dataset we are currently working with spans only a four-month period. To ensure the validity and robustness of our inferences, there is a need to aggregate data over a longer timeframe. One potential solution is to reacquire data from Reddit, covering a period of one year. However, this approach presents certain challenges, including the necessity for dimension reduction and the undertaking of extensive feature engineering to discern the most crucial variables that drive our analysis.

Techniques Applied

Clustering

One prominent technique we will employ is clustering. This technique will help us identify subtypes within the realm of depression. By clustering similar posts, we can unveil potential variations and commonalities across different aspects of depression (Kung, 2022). It's worth noting that there are several clustering methods that can be applied, such as density-based and K-means clustering. In our further analysis, we will compare the effectiveness of these models.

After trying k-means and DBSCAN clustering with some trial and error, the project ultimately settled on using t-SNE clustering. One of the primary reasons for selecting t-SNE is its superiority in handling complex datasets, like text data from subreddits, which often involve numerous variables and intricate relationships. Unlike traditional methods such as k-means or DBSCAN, t-SNE excels in revealing the underlying structure of the data by reducing dimensionality in a way that preserves local relationships between points. This makes it particularly adept at illustrating subtle groupings and patterns within large text datasets.

The application of t-SNE in our project aids in identifying nuanced clusters within the depression discussions. These clusters might represent different themes, tones, or contexts of discussions, providing valuable insights into the diverse expressions and experiences of depression in online communities. By leveraging t-SNE, we can effectively visualize and interpret these complex groupings, which would be challenging with other clustering methods.

Sentiment analysis

Sentiment analysis is another technique that could be applied in this study. By examining the sentiments expressed in posts, we can discern whether different clusters exhibit varying levels of depression. This approach will provide insights into the emotional states of individuals and help classify posts based on their associated sentiments. However, it's important to note that while sentiment analysis is valuable, we won't make it the main focus of our study. The reason for this is that we lack the real diagnoses of the authors of these posts, making it potentially invalid to use sentiment analysis to create a depression spectrum.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a crucial technique we have chosen for analyzing text data from depression-related subreddits. LDA is a type of probabilistic model that identifies topics present in a corpus of text. This model is particularly effective for our study as it helps uncover underlying thematic structures within the posts. Unlike clustering, which groups similar texts, LDA discerns the mixture of topics each post comprises, offering a more nuanced understanding of the content. This is especially pertinent in the context of depression, where discussions may involve a range of themes such as emotional states, coping mechanisms, personal experiences, and societal perceptions.

The decision to employ LDA is driven by its ability to handle large volumes of unstructured text data, like those found in subreddit posts. By analyzing these texts, LDA can reveal predominant topics and patterns that might not be immediately apparent. This insight is invaluable in understanding the complex nature of depression as discussed online. Moreover, LDA's topic modeling can potentially identify unique or emerging themes within these discussions, which may contribute to a deeper understanding of how depression is experienced and articulated by individuals. By leveraging LDA, we aim to add a layer of depth to our analysis, going beyond mere sentiment analysis to grasp the multifaceted nature of discussions surrounding depression.

Findings

The analysis has two phases. In the first phase, we look the data from December 2018 to October 2019. The second stage we merge the data with subreddits from January 2020 to April 2020.

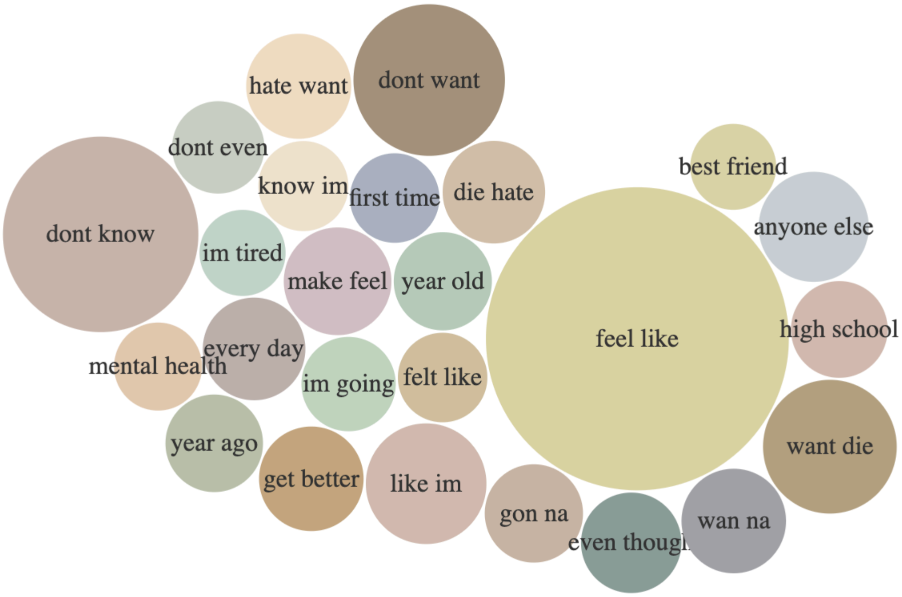

Figure.3 Frequent bi-grams bubble chart

The two word clouds indicate that individuals who have experienced depression tend to predominantly discuss their emotions when sharing their stories on online platforms. The feeling of loneliness is highly prevalent, and the word "friend" also occurs frequently. In the trigrams plot, people frequently express their disappointment with life. Another noteworthy pattern in the trigrams data is the frequent mention of the phrase "they don't know," indicating that individuals experiencing depression often struggle to understand the origins of their feelings. Additionally, suicide is mentioned as a high-frequency term, underscoring the critical importance of early intervention for individuals dealing with depression.

Index	Topics
0	depression, help, year, time, get, anxiety, feel, month, really, mental
1	life, day, work, get, thing, job, time, like, go, even
2	year, friend, school, got, time, one, would, back, get, since
3	feel, like, know, people, want, really, friend, thing, even, life
4	im, dont, want, cant, know, like, feel, life, fucking, even

Table.1 Topic modeling results

An intriguing observation from the analysis of the two phases, the shorter and the longer periods, is that the most frequent topics have remained relatively consistent. This consistency may reveal a persistent thematic core within the discussions related to depression, indicating that certain aspects of the experience maintain their significance over time. The recurring keywords such as 'depression', 'help', 'anxiety', and 'feel' across different times suggest that individuals continue to seek support and share similar struggles regardless of the period. Similarly, words like 'life', 'day', 'work', and 'school' imply that everyday life contexts remain central to the discourse around depression. This stability in topics may point to enduring patterns in how people communicate their experiences with depression, emphasizing the need for sustained attention to these themes in mental health support and interventions.

Clusters	Subtopics 1	Subtopics 2	Subtopics 3
1	gt, blog, john, st, wort, betterhelp, psilocybin	playlist, spotify, saddest, 21st, nd, rap, fortnite	depression, im, like, feel, get, know, life
2	im, dont, feel, like, know, want, cant	likethey, againi, outfit, powerfull, tok, everyond, verry	sam, inconsequential, deployment, soim, fase, aloneplease, terrorised
3	feel, like, know, want, dont, life, im	nighttime, comfy, singled, fluoxetine, 1am, tfw, freewill	distinct, akward, poop, strategy, despairdread, exausting, fingertip

Table.2 Clustering topic modeling results

Post-clustering topic modeling delineates diverse discussion threads within the depression narrative. Cluster 1 combines therapeutic discourse with cultural elements. Keywords such as "gt," "blog," "john," and "st. wort" indicate discussions around a spectrum of treatments, complemented by "playlist," "spotify," and "rap" which signal the use of music as a coping mechanism and its influence on users' mood and daily life.

Cluster 2's lexicon—"im," "dont," "feel"—portrays the emotional and psychological expression within the community. Unconventional terms like "likethey" and "tok" may signify unique personal narratives or cultural idioms. "Deployment" suggests conversations around profound life changes and their emotional toll.

Cluster 3 fixates on intimate sentiments and experiential accounts, with "nighttime," "comfy," and "fluoxetine" pointing to day-to-day realities and treatment discussions. "Distinct," "strategy," and "despair" likely relate to personal strategies for managing depression.

In essence, topic modeling reveals a rich dialogue spanning therapeutic choices, emotional well-being, and psychological strategies, informing nuanced support and intervention strategies.

Bibliography

1. Goodwin, R. D., Dierker, L. C., Wu, M., Galea, S., Hoven, C. W., & Weinberger, A. H. (2022). Trends in U.S. Depression Prevalence From 2015 to 2020: The Widening Treatment Gap. *American Journal of Preventive Medicine*, 63(5), 726–733. https://doi.org/10.1016/j.amepre.2022.05.014

2. 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), e22635–e22635. https://doi.org/10.2196/22635

3. Kung, B., Chiang, M., Perera, G., Pritchard, M., & Stewart, R. (2022). Unsupervised Machine Learning to Identify Depressive Subtypes. *Healthcare Informatics Research*, 28(3), 256–266. https://doi.org/10.4258/hir.2022.28.3.256

4. Tejaswini, V., Babu, K. S., & Sahoo, B. (2022). Depression Detection from Social Media Text Analysis using Natural Language Processing Techniques and Hybrid Deep Learning Model. *ACM Transactions on Asian and Low-Resource Language Information Processing*. https://doi.org/10.1145/3569580