Understanding COVID-19's Impact on Mental Health: A Reddit-Based Analysis

Mitul Vashista (B20AI022)
vashista.1@iitj.ac.in

Abstract

This study investigates the impact of the COVID-19 pandemic on mental health by analyzing user-generated content from Reddit. Utilizing natural language processing (NLP), topic modeling, emotion detection, and sentiment analysis, we evaluate trends and patterns across several mental health subreddits. Furthermore, we employ advanced machine learning techniques including SBERT embeddings, hierarchical clustering, UMAP, and random forest classification to uncover latent structures and predict subreddit origins of posts. The results show a general decline in sentiment over the course of the pandemic, dominant emotions such as sadness and trust, and important themes including emotional dysregulation and job insecurity.

1 Introduction

The COVID-19 pandemic has profoundly disrupted lives worldwide, raising concern over its psychological and emotional toll. Social media, especially platforms like Reddit, serve as real-time indicators of public sentiment. Reddit allows candid sharing due to its relative anonymity, making it a valuable source for mental health discourse. This project explores mental health-related discussions during COVID-19 using sentiment and emotion analysis, clustering, and machine learning to extract insights relevant to clinicians, researchers, and policymakers.

2 Dataset and Data Collection

Reddit posts were collected using the PRAW API across five mental health subreddits: `depression`, `Anxiety`, `mentalhealth`, `SuicideWatch`, and `bipolar`. Using the keyword "covid", approximately 5000 posts were extracted. Posts with less than 50 characters were filtered out. Each post includes subreddit name, post ID, date, and content. Data collection covered the period from early 2020 to early 2025.

3 Data Preprocessing

Python was used to clean the data by removing URLs, converting text to lowercase, and removing stopwords. Text tokenization and lemmatization were also applied. The resulting dataset was saved as `mental_health_covid_posts.csv`. Sample posts from each subreddit were manually reviewed for quality.

4 Methods

4.1 Sentiment and Emotion Analysis

Using the `syuzhet` package in R, sentiment scores were derived using three lexicons: Bing, Afinn, and Syuzhet. Emotional classification was performed using NRC Emotion Lexicon, yielding scores for emotions such as joy, sadness, trust, and fear. An illustrative post labeled with NRC: "I'm so tired of feeling anxious every day" yielded high scores in fear and sadness.

4.2 Topic Modeling

Latent Dirichlet Allocation (LDA) was used to extract dominant topics across subreddits, revealing recurring themes related to anxiety, job loss, emotional dysregulation, and medication effects. Number of topics per subreddit was determined using coherence scores.

4.3 SBERT Embeddings and Clustering

Text embeddings were generated using the `all-MiniLM-L6-v2` model from Sentence-Transformers. Dimensionality was reduced using UMAP and clustering was performed using Agglomerative Clustering (Ward linkage). Five clusters emerged representing underlying themes in discussions. A thematic summary of the clusters is shown in Table 1.

Table 1: Summary of UMAP clusters

Cluster ID	Dominant Theme	Top Keywords	Sample Phrase
0	Anxiety and Fear	anxious, panic, overwhelmed	"I can't sleep due to constant panic attacks"
1	Medication and Side Effects	meds, dosage, side effects	"This new prescription is making me dizzy"
2	Isolation and Loneliness	alone, isolated, empty	"Haven't seen anyone in weeks"
3	Suicidal Ideation	ending, hopeless, tired	"I don't want to live anymore"
4	Bipolar Episodes	mania, mood, swings	"I was manic for two days straight"

4.4 Classification

A Random Forest classifier was trained to predict the subreddit origin of posts using SBERT embeddings. The model achieved 63% accuracy, with class-wise precision ranging from 44% (`depression`) to 88% (`bipolar`). Error analysis showed confusion between `depression` and `mentalhealth`, likely due to thematic overlap.

5 Results and Visualization

5.1 Sentiment Trends by Subreddit

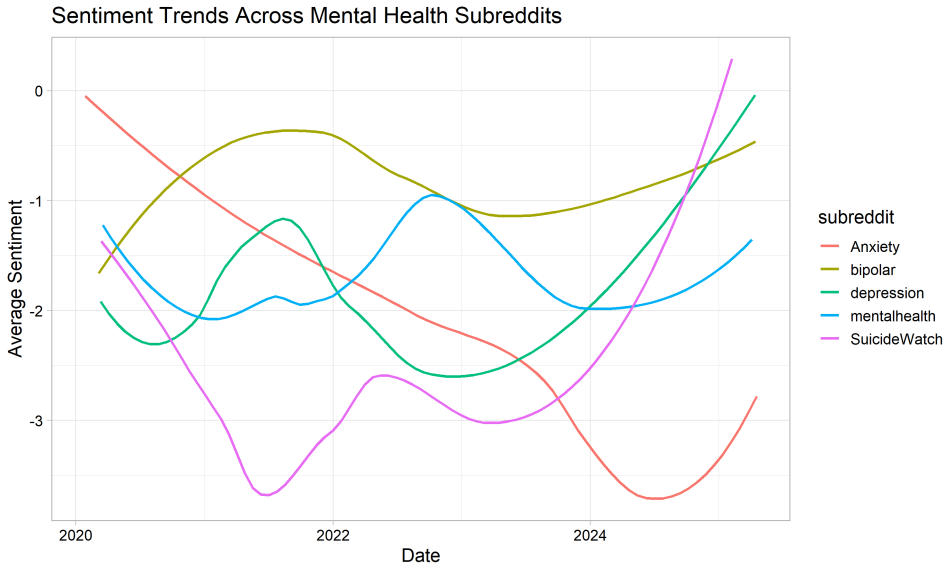

Figure 1: Sentiment trends over time across subreddits. Sentiment declined initially during the pandemic and showed signs of recovery post-2024.

5.2 Emotion Distribution by Subreddit

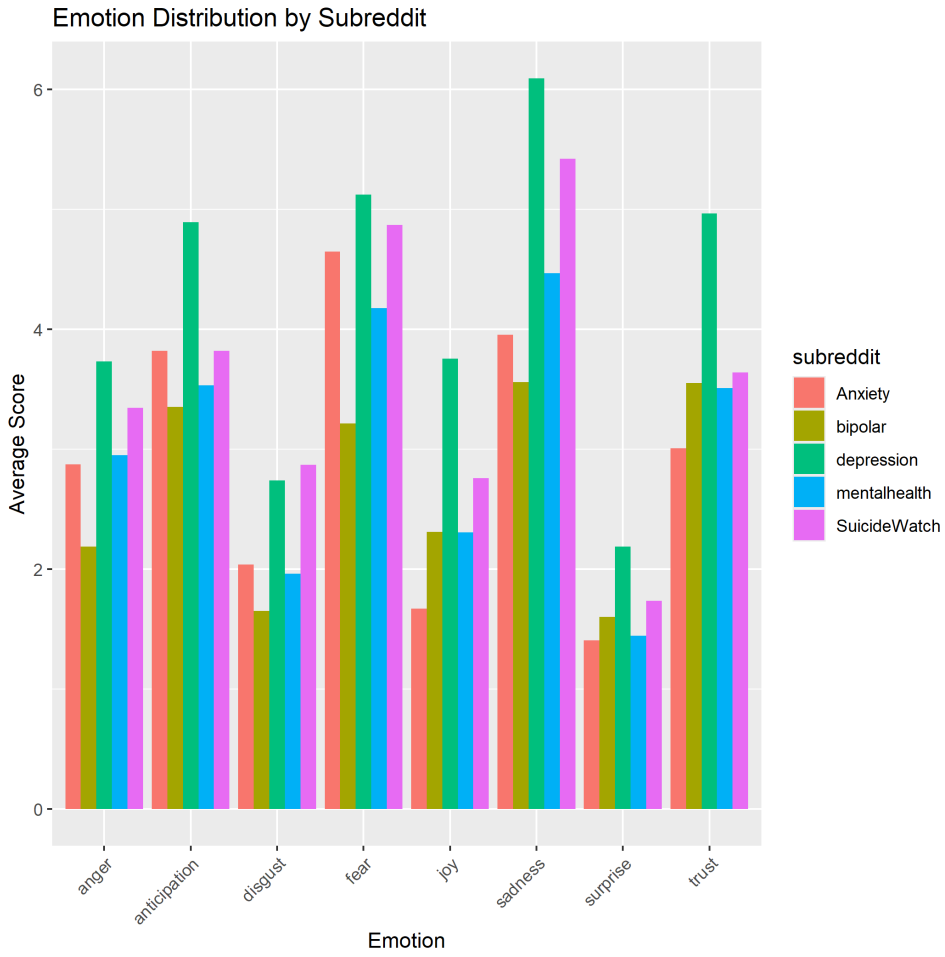

Figure 2: Distribution of emotions across subreddits. Sadness, trust, and anticipation are consistently dominant.

5.3 UMAP and Clustering

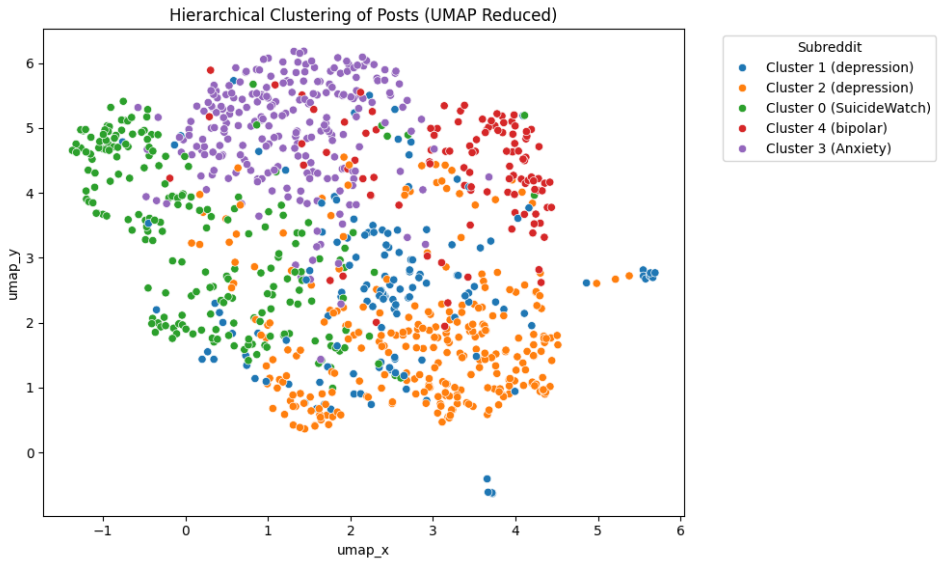

Figure 3: UMAP-based clustering of SBERT-encoded posts.

5.4 Word Clouds

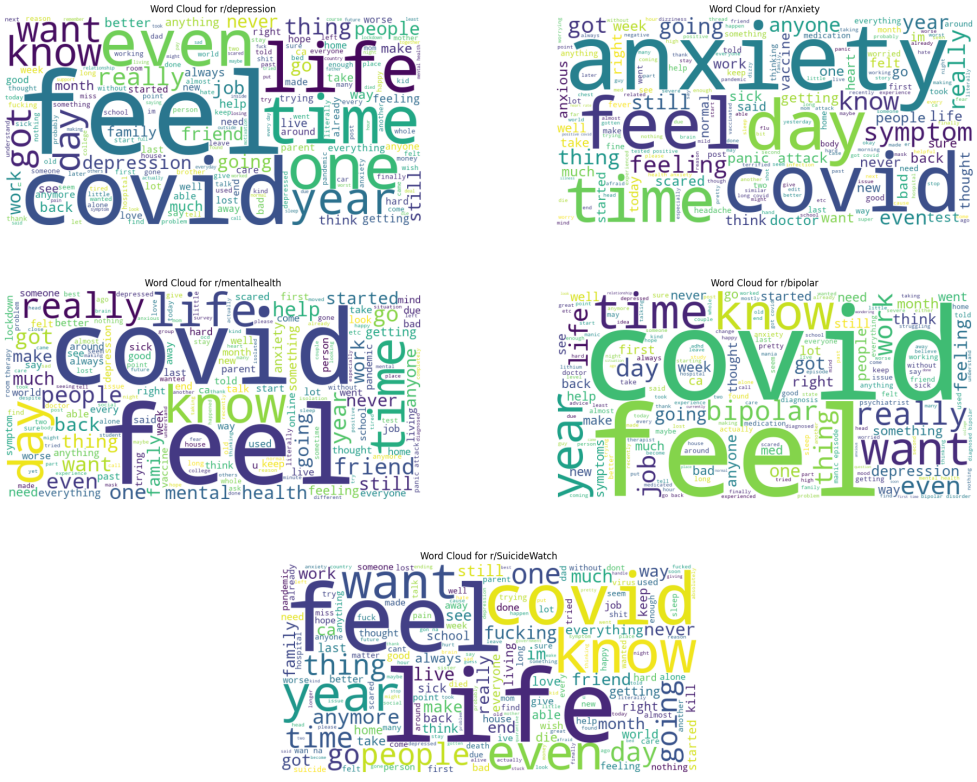

Figure 4: Word clouds for individual subreddits.

5.5 Classification Results

Table 2: Random Forest classification report

Subreddit	Precision	Recall	F1-score
r/depression	0.44	0.42	0.43
r/Anxiety	0.60	0.57	0.58
r/mentalhealth	0.66	0.70	0.68
r/SuicideWatch	0.55	0.50	0.52
r/bipolar	0.88	0.85	0.86

6 Discussion

Our results reveal key emotional and thematic patterns during the pandemic:

- Declining sentiment in early pandemic stages with gradual recovery post-2024.
- High levels of sadness, trust, and anticipation, indicating both distress and hope.
- SBERT clustering exposed nuanced post groupings beyond subreddit labels.
- Word clouds indicate subreddit-specific concerns (e.g., "mania" in `bipolar`, "lonely" in `SuicideWatch`).
- Misclassifications in subreddit prediction reflect overlapping psychological concerns across communities.

7 Conclusion

This study demonstrates the utility of Reddit as a source for real-time mental health surveillance. By integrating NLP, statistical modeling, and machine learning, we extract actionable insights that can inform targeted intervention and policy. Our findings highlight the psychological burden carried by individuals throughout the COVID-19 pandemic, as evidenced by dominant emotions such as sadness and anticipation, and topics including medication effects, loneliness, and job insecurity. The observed gradual improvement in sentiment post-2024 offers cautious optimism, but also underscores the lingering emotional consequences of the pandemic.

This analysis supports the potential of social media platforms as a means of understanding mental health at scale and in real-time. It can complement traditional survey-based research and provide timely signals for healthcare professionals and policymakers aiming to mitigate the effects of future global crises.

8 Future Work

- Investigate how COVID-19 acts as a trigger for psychiatric symptoms, even in individuals without prior mental health history.
- Examine the bidirectional relationship between COVID-19 and mental health, with a focus on how pre-existing conditions may exacerbate infection risk and outcomes.
- Study the neurological impacts of SARS-CoV-2, including its influence on cognitive and emotional processing.
- Use more sophisticated NLP models such as BERT-based sentiment and emotion classifiers to improve classification granularity.
- Extend the dataset to include global subreddits and other platforms like Twitter and health forums for cross-platform analysis.
- Perform longitudinal tracking of post-pandemic mental health trajectories beyond 2025.
- Benchmark the current Random Forest model against other classifiers (e.g., SVM, Naive Bayes) and deep learning approaches.
- Incorporate network analysis to understand how users interact and support each other within mental health communities.
- Conduct comparative studies across different mental health subreddits to identify unique versus shared stressors and support mechanisms.

9 References

- Syuzhet: `https://github.com/mjockers/syuzhet`
- PRAW API: `https://praw.readthedocs.io/en/latest/`
- Sentence Transformers: `https://www.sbert.net/`
- NRC Emotion Lexicon: `https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`