Social Media Text Analysis for Early Detection of Mental Health Symptoms Final Report

Vaibhav Rawat
a1914731

November 25, 2025

Report submitted for **MATHS 7097B - Data Science Research Project B** at the School of Mathematical Sciences, University of Adelaide

THE UNIVERSITY of ADELAIDE

Project Area: **Natural Language Processing**
Project Supervisor: **Dr. INDU BALA**

Abstract

Digital traces on social media give us a unique look at how people talk about and deal with their mental health in their daily lives. This project creates a multi-layered analytical framework for Reddit posts from mental health-related communities. It does this by combining emotion classification based on transformers, mapping to the Circumplex Model of Affect, and network modeling of communities and emotional transitions. We used the Python Reddit API Wrapper (PRAW) to get data from a carefully chosen group of mental health and support subreddits. Each subreddit could only have 1,000 posts, and only posts from 2023 and later were allowed. After a lot of cleaning and exploratory analysis, each post went through the cardiffnlp/twitter-roberta-base-emotion The RoBERTa model gives us probabilistic scores for four main emotions: happiness, hope, anger, and sadness. We used a weighted mapping based on the Circumplex Model of Affect to turn these probabilities into continuous valence and arousal coordinates. This gave each post a two-dimensional affective representation.

Next, the analysis looked at the structure of relationships instead of just individual posts. A subreddit co-occurrence network was created by connecting communities that had active users in common. A directed emotion transition network, on the other hand, showed how the most common emotions changed in discussion threads. Lastly, an interactive Streamlit dashboard was created to let people explore these different levels of analysis. Users can filter by subreddit and look at emotion distributions, valence-arousal scatters, and example posts. The results indicate that the majority of posts reside in a low-valence, moderate-to-high-arousal quadrant of affective space, characterized by prevalent sadness across numerous communities, albeit with significant variation between diagnostic forums and supportive, recovery-oriented subreddits. The subreddit co-occurrence network shows that there are very connected hubs, like general mental health and depression communities, that are in between more specialized spaces. The emotion transition network shows that negative feelings can last for a long time and that a significant number of people move from sadness to optimism or joy after getting supportive replies. The project shows how using transformer models, affective theory, and network analysis together can help us better keep an eye on online mental health discussions. It also gives us a starting point for more clinically focused early-warning tools.

1 Introduction

Despite the fact that mental health diseases account for a considerable portion of the world's disease burden, the majority of individuals who suffer from severe psychological discomfort never seek out professional psychiatric services. At the same time, millions of people use internet forums to share their problems, look for support, and connect with others. Particularly on Reddit, there is a thriving community of people discussing mental health issues, including symptoms, diagnosis, treatment experiences, crises, and daily coping. Because these discussions are pseudonymous, extremely textual, and chronologically rich, Reddit is a useful but intricate data source for learning how interaction patterns and emotional expressions change in realistic environments.

Static prediction tasks, including determining whether a person has depression or suicide risk based on their postings or assigning emotion labels to individual messages, have accounted for a large portion of computational work on mental health and social media. These methods are useful, but they frequently ignore the larger emotional dynamics of conversations and groups and consider posts as separate entities. It might be more instructive to describe how emotions move inside threads and across networked groups, as well as how various online places display unique "emotional climates," in order to detect problems early and take appropriate action.

The goal of this study is to help bring about that change by creating a framework that examines how Reddit communities' emotional expressions and interaction patterns might be early indicators of mental health issues. Three methodological layers are integrated instead of depending only on basic sentiment labeling. Initially, each post's nuanced emotion probabilities are obtained using transformer-based models. In order to characterize posts in terms of both positivity/negativity and activation, we first use the Circumplex Model of Affect to map these emotion scores into a continuous valence–arousal space. Third, network models are utilized to capture the way emotions change within threads and how subreddits relate to one another through shared users.

The main research topics are: How are emotions distributed among Reddit's various mental health subreddits? In what ways do posts from these communities fill the valence–arousal space, and do specific subreddit kinds occupy different areas? In what ways do communities share people, and what emotional transition patterns show themselves among threads? By providing answers to these questions, the project hopes to create a multi-scale image of online conversation around mental health that can eventually guide monitoring and early detection tactics.

The remainder of this study initially provides a brief overview of relevant research on network perspectives, transformer-based emotion detection, social media mental health analysis, and the Circumplex Model. The techniques employed are then thoroughly explained, including data collection, preprocessing, valence-arousal mapping, RoBERTa-based emotion rating, subreddit and emotion networks, and the interactive dashboard. Each layer of analysis's results are shown, and then their ramifications, limits, and future research directions are discussed.

2 Background

Over the past ten years, research on mental health and social media has grown significantly. Early research aimed to detect signs of depression or suicidal thoughts at the population level by analyzing the sentiment and keywords of social media sites like Twitter. Later study focused on user-level categorization, using linguistic cues, posting habits, and metadata to train algorithms to differentiate self-identified depressed users from controls. In forums like r/depression, r/SuicideWatch, r/Anxiety, and r/BipolarReddit, many members openly share their diagnoses or mental health conditions, which has led to Reddit's growing popularity. This has made it possible to analyze symptom language, help-seeking, and support dynamics in peer-to-peer settings more precisely.

At the same time, transformer architectures like BERT and RoBERTa have revolutionized the larger field of natural language processing. These models, which have been pre-trained on large corpora and optimized for certain tasks, routinely surpass earlier methods in sentiment categorization and emotion recognition. In mental health applications, specialized variations like MentalBERT and ClinicalBERT have demonstrated enhanced sensitivity to clinical terminology and subtle emotional cues after being trained on counseling notes, clinical narratives, or mental health-related social media corpora. Some of these specialized models, however, have gated access, and tweaking them necessitates tagged data, which are frequently hard to come by.

2.1 Related Work

The majority of emotional natural language processing (NLP) work still depicts emotion using coarse polarity (positive/negative) or discrete categories (such as joy, sadness, or rage). However, the literature on psychology has traditionally placed a strong focus on continuous models of affect. According to the Circumplex Model of Affect, valence (pleasant–unpleasant) and arousal (activated–deactivated) characterize emotions in a two-dimensional space. Anger and enthusiasm are both high arousal emotions, however they have different valences, whereas melancholy is negative and low arousal. More detailed descriptions of emotional states can be achieved by mapping text into this area. Previous research has estimated the valence and arousal of words, phrases, or messages using regression or lexicon-based methods.

Network viewpoints offer an additional perspective on mental health. While social-network analyses look at how relationships and interactions affect mental health outcomes, symptom network models see diseases as interacting systems of symptoms rather than signs of a hidden disease.

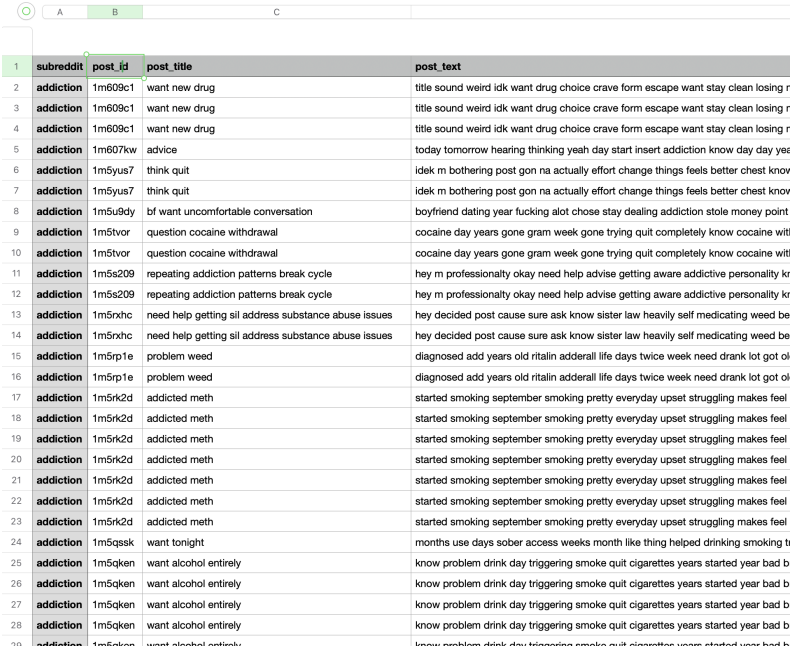
On sites such as Reddit, individuals engage with various communities, and discussions develop as a series of posts and comments. By modeling these relationship structures, it is possible to see how supportive reactions are dispersed, how distress spreads, and which communities are crucial to users' trajectories.

Despite advancements, comparatively few research apply these three strands—continuous affect spaces, network analysis, and transformer-based emotion modeling—to mental health communities in a unified framework. By creating a pipeline from raw Reddit postings to interconnected perspectives of emotion, affect, and network structure, this project aims to close that gap. It places a special emphasis on scalable techniques that can handle big, noisy datasets.

3 Methods

3.1 Data Collection

This study gathered posts and comments from 36 subreddits that address psychological conditions, emotional well-being, support networks, and recovery experiences in order to create a thorough and varied dataset of mental health discourse. These subreddits were chosen because they are relevant to a wide range of mental health conditions, such as mood disorders (e.g., bipolar disorder, depression), anxiety disorders (e.g., anxiety, social anxiety, PTSD), eating disorders (e.g., anorexia nervosa, binge eating disorder), substance use (addiction, cease drinking), neurodevelopmental disorders (ADHD, AutismInWomen), and online mental health support groups (e.g., KindVoice, SuicideWatch).



	A	B	C	
1	subreddit	post_id	post_title	post_text
2	addiction	1m609c1	want new drug	title sound weird idk want drug choice crave form escape want stay clean losing r
3	addiction	1m609c1	want new drug	title sound weird idk want drug choice crave form escape want stay clean losing r
4	addiction	1m609c1	want new drug	title sound weird idk want drug choice crave form escape want stay clean losing r
5	addiction	1m607kw	advice	today tomorrow hearing thinking yeah day start insert addiction know day day yea
6	addiction	1m5yus7	think quit	idek m bothering post gon na actually effort change things feels better chest knov
7	addiction	1m5yus7	think quit	idek m bothering post gon na actually effort change things feels better chest knov
8	addiction	1m5u9dy	bf want uncomfortable conversation	boyfriend dating year fucking alot chose stay dealing addiction stole money point
9	addiction	1m5tvor	question cocaine withdrawal	cocaine day years gone gram week gone trying quit completely know cocaine with
10	addiction	1m5tvor	question cocaine withdrawal	cocaine day years gone gram week gone trying quit completely know cocaine with
11	addiction	1m5s209	repeating addiction patterns break cycle	hey m professionalty okay need help advise getting aware addictive personality kr
12	addiction	1m5s209	repeating addiction patterns break cycle	hey m professionalty okay need help advise getting aware addictive personality kr
13	addiction	1m5rxhc	need help getting sil address substance abuse issues	hey decided post cause sure ask know sister law heavily self medicating weed be
14	addiction	1m5rxhc	need help getting sil address substance abuse issues	hey decided post cause sure ask know sister law heavily self medicating weed be
15	addiction	1m5rp1e	problem weed	diagnosed add years old ritalin adderall life days twice week need drank lot got ol
16	addiction	1m5rp1e	problem weed	diagnosed add years old ritalin adderall life days twice week need drank lot got ol
17	addiction	1m5rk2d	addicted meth	started smoking september smoking pretty everyday upset struggling makes feel
18	addiction	1m5rk2d	addicted meth	started smoking september smoking pretty everyday upset struggling makes feel
19	addiction	1m5rk2d	addicted meth	started smoking september smoking pretty everyday upset struggling makes feel
20	addiction	1m5rk2d	addicted meth	started smoking september smoking pretty everyday upset struggling makes feel
21	addiction	1m5rk2d	addicted meth	started smoking september smoking pretty everyday upset struggling makes feel
22	addiction	1m5rk2d	addicted meth	started smoking september smoking pretty everyday upset struggling makes feel
23	addiction	1m5rk2d	addicted meth	started smoking september smoking pretty everyday upset struggling makes feel
24	addiction	1m5qssk	want tonight	months use days sober access weeks month like thing helped drinking smoking t
25	addiction	1m5qken	want alcohol entirely	know problem drink day triggering smoke quit cigarettes years started year bad b
26	addiction	1m5qken	want alcohol entirely	know problem drink day triggering smoke quit cigarettes years started year bad b
27	addiction	1m5qken	want alcohol entirely	know problem drink day triggering smoke quit cigarettes years started year bad b
28	addiction	1m5qken	want alcohol entirely	know problem drink day triggering smoke quit cigarettes years started year bad b

Figure 1: Collected Data

Since these platforms frequently feature emotionally charged narratives, confessions, and support-seeking behavior that reflect psychological distress even in the absence of formal diagnoses, it was also deliberate to include subreddits like offmychest, TrueOffMyChest, and needafriend. It is ideally suited for subsequent tasks like symptom severity analysis, emotion mapping, and network modeling because of its broad selection, which guarantees both depth and diversity in linguistic patterns, emotional expression, and community interaction.

3.2 Pre-processing and EDA

Because it includes mistakes, emojis, hyperlinks, formatting artifacts, and unique markup, Reddit content is intentionally noisy. Pandas, nltk, BeautifulSoup, and regular expressions were used to create a comprehensive cleaning pipeline in Python that prepared the data for transformer-based modeling as well as human-interpretable exploratory visualisations.

Eliminating entries with blank or missing material was the first stage. Posts lacking a title and body were eliminated, as were comments that included only whitespace or markers that had been erased. This made sure that every entry in the dataset had meaningful text that the models could use.

To lessen sparsity caused by case variations, all remaining text fields, including post titles, body text, and comment bodies, were changed to lowercase. Emojis and other non-standard Unicode symbols were eliminated using regular-expression patterns, while HTML elements and Markdown artifacts were eliminated using BeautifulSoup. For modeling purposes, URLs were either eliminated completely or replaced with a placeholder token because raw links and shorteners are usually not indicative of mental state. While it was kept for the transformer models, which can take advantage of sub-word structure and punctuation cues, it was eliminated for exploratory token-frequency analysis.

English stopwords were eliminated using the standard stopword list, and the cleaned text was tokenized using NLTK's word tokenizer for EDA. This resulted in token sequences that were utilized to create n-gram statistics, word clouds, and word-frequency tables. The distribution of post durations was depicted by exploratory plots, which revealed that there was a significant tail of lengthier narratives and venting postings but that many posts were brief disclosures or queries. Basic descriptive statistics were calculated for scores and comment counts as approximate indicators of community engagement, and subreddit-level counts revealed which communities produced the most content.

Each post's title and body content were concatenated with a separating period to form a full_text variable for downstream modeling. This made sure that the transformer model received all of the context that was accessible at the submission level in a single input sequence. The cleaned dataset, which served as the foundation for network creation and emotion categorization, was saved as cleaned_reddit_data.csv.

3.3 Classification of emotions using RoBERTa

The pipeline's next step was to give each post a complex emotional profile. Since domain-specific models like MentalBERT are trained on in-

formation relevant to mental health and may be able to catch delicate symptom language and slang, they were initially taken into consideration. Practical constraints, however, emerged: the MentalBERT model utilized in some earlier work is housed in a gated repository on Hugging Face, necessitating specific access authorization and authentication. Mental/mental-bert-base-uncased was not suited for simple, reproducible use in this project because attempts to load it produced authentication issues. Rather, the project used the cardiffnlp/twitter-roberta-base-

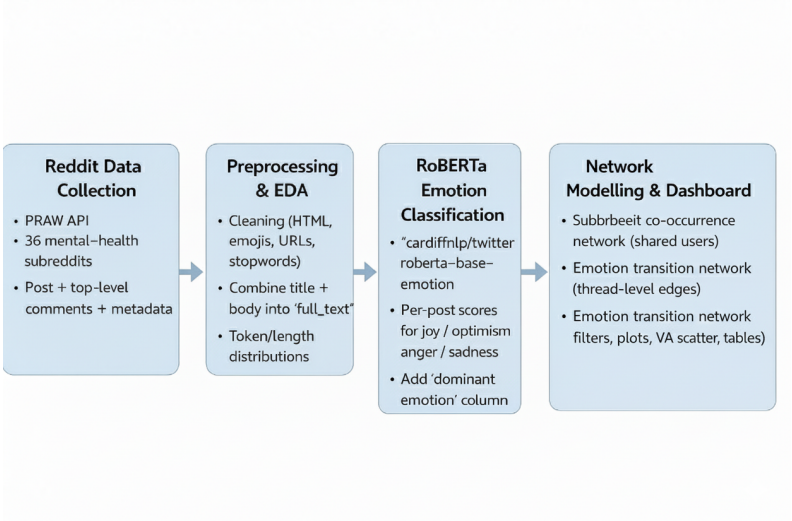


Figure 2: Process Pipeline

emotion model, which is freely available. This model, which was refined on a sizable Twitter emotion dataset and is based on the RoBERTa architecture, generates probability distributions across four major emotion categories: happiness, optimism, anger, and sadness. Despite not being trained on Reddit or mental health data specifically, it is a plausible option for a first-pass emotional analysis because it is reliable, well-documented, and frequently used in affective computing research.

The Hugging Face Transformers library was utilized in the implementation. After loading AutoTokenizer.from_pretrained and AutoModelForSequenceClassification.from_pretrained onto the model and tokenizer, a text-classification pipeline was applied. In order to acquire a full probability vector for each of the four emotions while adhering to the model's maximum sequence length, posts were fed into this pipeline in batches of 32 via a loop over the full_text column with return_all_scores=True and truncation=True enabled. Processing was tracked using a tqdm progress bar. In order to prevent the pipeline from breaking, basic error handling was put in place. If the model received an unexpected input and reported

an error in a batch, the script logged the batch index and replaced those items with a neutral, zeroed score vector.

For every post, the pipeline generated a dictionary that mapped emotion labels to probabilities that added up to one. To create a new dataset called reddit_emotion_roberta_full.csv, they were put together into a pandas DataFrame and concatenated with the important contextual columns from the cleaned data: subreddit, post ID, title, and text. The idxmax over the emotion scores was used to calculate a dominant_emotion column in addition to the four likelihood columns. While the entire probability distribution preserves information about mixed emotional states, this offers a straightforward categorical label that can be utilized for summary statistics. Initial analysis revealed trends that were in line with intuition: postings in more solution-focused or supportive communities frequently displayed a mix of melancholy and optimism, whereas posts in severely disturbed subreddits frequently had high sadness scores close to one.

3.4 Mapping to the Circumplex Model of Affect

Although distinct emotion categories are helpful, they fall short in describing mixed feelings or variances in intensity. The Circumplex Model of Affect, which arranges emotions along two axes—valence, which represents pleasantness vs unpleasantness, and arousal, which represents activation versus deactivation was used in the project to give a continuous affective representation. By mapping every post into this area, we may determine whether a text is simply "sad" or, for instance, low arousal and resigned as opposed to high arousal and agitated.

A straightforward but understandable mapping was created because the RoBERTa model produces probabilities for four emotions instead of raw valence and arousal scores. Based on psychological theory and intuition, coordinates in the valence–arousal space were manually assigned to each of the four emotion categories. Anger was positioned at strongly negative valence but high arousal $(-0.8,\ 0.8)$; sadness at very negative valence but substantially lower arousal $(-0.9,\ 0.3)$; optimism at moderately high valence and moderate arousal $(0.7,\ 0.5)$; and joy at high positive valence and high arousal $(0.9,\ 0.7)$. These positions represent the notion that optimism is positive but frequently calmer than strong joy, and that anger is an engaged form of negative affect, whereas sadness is more deactivated and withdrawn.

For each post i, let $p_{i,e}$ denote the probability assigned by RoBERTa to emotion e, and let (v_e, a_e) denote the valence and arousal coordinates of emotion e. The post-level valence V_i and arousal A_i were computed

as weighted sums:

$$V_i = \sum_e p_{i,e}\, v_e, \qquad A_i = \sum_e p_{i,e}\, a_e. \tag{1}$$

$$\text{Valence} = p_{\text{joy}} \cdot v_{\text{joy}} + p_{\text{optimism}} \cdot v_{\text{optimism}} + p_{\text{anger}} \cdot v_{\text{anger}} + p_{\text{sadness}} \cdot v_{\text{sadness}}$$

$$\text{Arousal} = p_{\text{joy}} \cdot a_{\text{joy}} + p_{\text{optimism}} \cdot a_{\text{optimism}} + p_{\text{anger}} \cdot a_{\text{anger}} + p_{\text{sadness}} \cdot a_{\text{sadness}}$$

$$\text{Valence} = 0.9\, p_{\text{joy}} + 0.7\, p_{\text{optimism}} - 0.8\, p_{\text{anger}} - 0.9\, p_{\text{sadness}}$$

$$\text{Arousal} = 0.7\, p_{\text{joy}} + 0.5\, p_{\text{optimism}} + 0.8\, p_{\text{anger}} + 0.3\, p_{\text{sadness}}$$

Although this mapping is heuristic and only uses a limited number of emotions, it offers a useful method of utilizing the transformer's output while adhering to accepted emotional theory. Additionally, it yields results that are intuitive: postings with a high probability of melancholy end up close to the low-valence, somewhat low-arousal region, posts with mixed sadness and rage sit in a more activated negative area, and those that are joyous or hopeful inhabit the positive side of the plane.

3.5 Network Modelling

In addition to per-post affect, the project aimed to comprehend relational structure on two levels: how emotions shift within threads and how communities are connected by shared users.

The cleaned dataset with subreddit names, post authors, and comment authors was used for the subreddit-level network. Building a co-occurrence graph with each node standing for a subreddit and edges signifying that at least one user has participated in both communities was the aim. This was created by combining all instances of post_author and comment_author into a single "user" column while maintaining the related subreddit. The number of different subreddits that each user participated in was calculated. The unordered pairs of those subreddits were then created using combinations for each user who was a member of at least two subreddits. Between each pair, an undirected edge was added, with an edge weight that takes into account the number of users that the two communities share. The graph structure was managed via NetworkX.

$$w_{ij} = \sum_u \mathbf{1}\big[u \in S_i \wedge u \in S_j\big], \qquad d_i = \sum_j w_{ij}$$

w_{ij} is the weight of the edge between subreddit i and subreddit j; it counts how many users u participate in *both* subreddits (i.e., their co-occurrence).

d_i is the degree of subreddit i, computed as the sum of its edge weights to all other subreddits:

$$d_i = \sum_j w_{ij}.$$

A larger value of d_i indicates that subreddit i is more strongly connected to the rest of the network. To highlight the interrelated aspect of the ecosystem, isolated nodes subreddits without any shared users were eliminated. For additional study, the largest connected component which stands for the center of the mental-health network was taken out. A straightforward indicator of centrality was node degree, or the number of neighbors, and node sizes in plots were scaled in accordance with degree. To draw attention to significant overlaps, edge widths were adjusted by co-occurrence weight. To distribute the nodes uniformly while maintaining apparent edge lengths, the Kamada–Kawai layout algorithm was selected.

The emotion-scored dataset reddit_emotion_roberta_full.csv was utilized for the thread-level emotion transition network. As previously mentioned, a dominant_emotion label was given to each row. In order to approximate threads, posts were grouped by post_id. Within each group, rows were arranged according to how they appeared in the dataset; timestamps may be used to further refine this order, but in reality, posts and comments for a certain ID are typically preserved in chronological order. Transitions from one contribution's dominating emotion to the next's dominant emotion were captured via an iterative scan for every thread. A directed edge from the prior emotion to the current one was put to a graph and its weight increased each time the emotion changed. The resulting compact four-node directed graph summarizes the tendency of emotions to flow across responses.

A spring layout was used to visualize the resulting directed graph. For clarity, node sizes were maintained constant, and edge widths and colors were mapped to the frequency of transitions: the most frequent transitions were shown as thick green edges, medium-frequency transitions as orange, and infrequently observed transitions as thin grey arcs. Readers could observe, for example, how frequently melancholy endures vs changing to optimism following a response thanks to edge labels that displayed precise counts.

3.6 Interactive Dashboard

An interactive dashboard was created with Streamlit to make the analyses easily accessible and enable flexible exploration. The valence-arousal dataset and the emotion-scored dataset are combined into a single table on the dashboard's backend. Potential duplicates were eliminated prior to merging because the valence–arousal file only has one row per post ID. A left join on post_id then made sure that each element in the emotion file received valence and arousal values where they were available. During load time, a convenience column called dominant_emotion was recalculated.

In the interactive dashboard, several summary quantities are computed directly from the model outputs.

First, for each subreddit s and emotion e, we compute the average emotion score over all posts in that subreddit:

$$\bar{E}_{s,e} = \frac{1}{N_s} \sum_{i \in \mathcal{P}_s} p_{i,e},$$

where $p_{i,e}$ is the RoBERTa probability for emotion e on post i, $\mathcal{P}_s$ is the set of posts in subreddit s, and $N_s = |\mathcal{P}_s|$.

For each post i, the dominant emotion used in the bar charts and example tables is defined as

$$\mathrm{dom}(i) = \arg \max_{e \in \{\text{joy, optimism, anger, sadness}\}} p_{i,e}.$$

The dashboard also displays global (or filtered) mean valence and arousal across the currently selected posts:

$$\bar{V} = \frac{1}{N} \sum_{i=1}^{N} V_i, \qquad \bar{A} = \frac{1}{N} \sum_{i=1}^{N} A_i,$$

where V_i and A_i are the valence and arousal scores for post i and N is the number of posts in the filtered subset.

Finally, the proportion of posts whose dominant emotion is e (used in the "Dominant Emotion Distribution" plot) is given by

$$\pi_e = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\mathrm{dom}(i) = e\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, equal to 1 if the condition is true and 0 otherwise. The dashboard is divided into a primary visualization section and a filtering sidebar. A multi-select widget in the sidebar allows users to choose one or more subreddits; by default, a number of

well-known communities are displayed. To give context, the number of posts in the current view is shown. Total posts, the number of distinct subreddits, and the mean valence and arousal of the filtered subset are displayed in the top row of metrics in the main area.

The emotional landscape is summarized by several plots beneath the metrics. The most troubled areas are highlighted in a bar chart that displays average emotion scores by subreddit for the chosen communities. Subreddits are arranged according to their average level of sadness. It is simple to see whether melancholy, anger, joy, or optimism are more prominent in the present view thanks to a second bar chart that shows the distribution of dominant emotions in the filtered set.

Each post is positioned in valence–arousal space using a scatter plot to visualize the Circumplex mapping. A random subsample is taken for performance if more than 5,000 postings are chosen. A "coolwarm" color map is used to color the points according to valence, and the plane is divided into quadrants by reference lines at zero valence and zero arousal. The distributions of the four emotion probabilities across the filtered posts are displayed next to this in a kernel-density estimate (KDE) plot, which gives an idea of how strongly each emotion tends to be exhibited. A customisable table of sample posts that shows the subreddit, title,

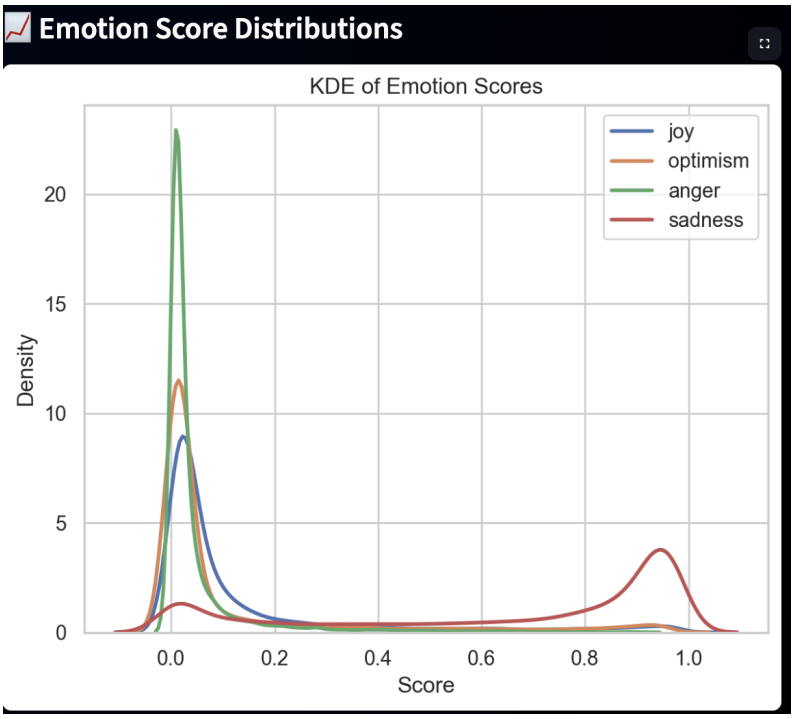


Figure 3: kernel-density estimate (KDE) plot

text, dominant emotion, emotion probabilities, and affective coordinates

follows a table summarizing subreddits by their mean emotion scores and valence-arousal coordinates at the bottom of the dashboard. These sample postings enable qualitative inspection to confirm that content that is clearly upset is associated with high sadness or rage scores. As a result, the dashboard integrates the previous analytical stages into an exploratory tool that practitioners or researchers may utilize.

4 Results

4.1 Emotional landscape from RoBERTa

Every post has a probabilistic emotion profile generated by the RoBERTa-based emotion classification. Overall, there is a significant bias towards sorrow in the distribution of dominating emotions, which is to be expected considering the emphasis on mental health issues. The likelihood of unhappiness is very high in many posts, which express clear feelings of being worn out, depressed, or incapable of handling life. Nonetheless, a significant percentage of posts also exhibit optimism and delight as the predominant feelings, especially in communities focused on sharing positive updates, offering support to others, or commemorating recovery milestones. Posts that indicate dissatisfaction with healthcare systems, employers, family members, or one's own perceived lack of progress are often the source of anger.

These trends are further illustrated by kernel-density plots of emotion scores for every post. Many posts are viewed by RoBERTa as nearly exclusively depressing, as evidenced by the right-skewed distribution of sadness scores with a significant mass near one. A subset of postings that are very optimistic or forward-thinking are highlighted by optimism scores, which peak closer to zero but have a lengthy right tail. The fact that even encouraging news in mental health communities is sometimes overshadowed by persistent difficulties is reflected in the generally lower joy scores. Although anger scores are generally low, there are noticeable surges in particular subreddits where venting and righteous outrage are prevalent.

These conclusions are supported by correlation analysis between the four emotion scores. There is a large negative correlation between joy and melancholy, meaning that posts tend to tilt in one direction rather than expressing both at high intensity. Though not as strongly, sadness and optimism are inversely correlated, which supports the notion that some posts may include both declarations of resolve or hope as well as feelings of grief. In some situations, anger and melancholy have a moderately positive link, which may be related to posts that discuss both pain and resentment. Despite being trained on Twitter instead of Reddit, these patterns demonstrate that the RoBERTa model is collecting meaningful structure rather than random noise.

Different emotional profiles can be found by averaging emotion scores by subreddit. Compared to environments that are crisis- or diagnosis-focused, support and encouragement societies exhibit comparatively higher levels of optimism and lower levels of despair. A mixture of melancholy, reflecting struggle and relapse, and noteworthy optimism, when people

report progress or express gratitude for support, can be seen in addiction recovery forums. More complex emotional combinations can be seen in autism-related communities; posts celebrating identity and community can show joy and optimism, but discussions of prejudice or fatigue can show despair and rage.

4.2 Valence-Arousal Distributions

The Circumplex transformation creates a thick cloud of points that represent the dataset's overall affective environment when posts are mapped into the valence–arousal space. With numerous dots grouped between moderate and extremely negative valence values, the global scatter plot clearly demonstrates a concentration in the negative valence side of the plane, as predicted. Different posts have different levels of arousal; some are low-arousal, showing numbness, exhaustion, or resigned melancholy, while others are high-arousal, reflecting terrified anxiety, furious rage, or urgent emergencies.

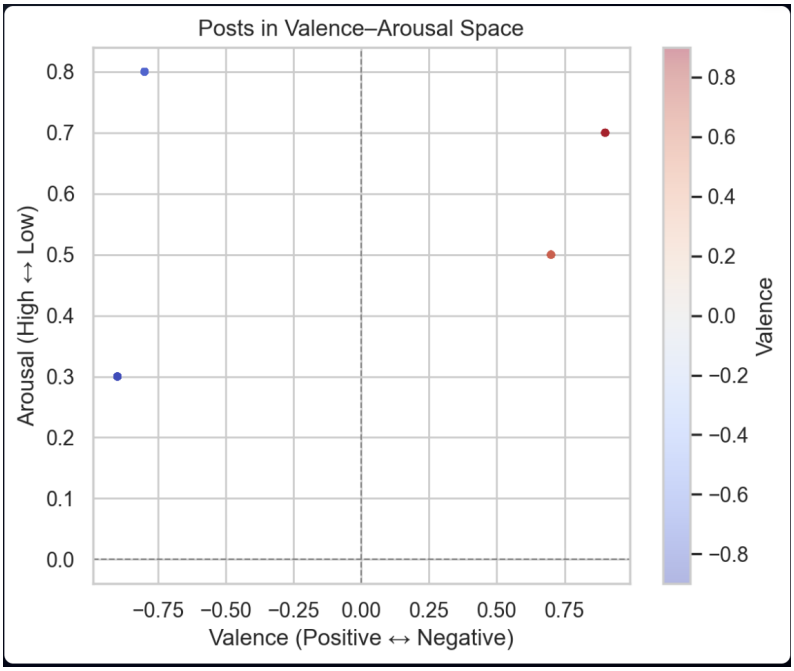


Figure 4: Valence-Arousal Distributions

There is a comparatively small group of points in the top-right quadrant (high valence, high arousal) that indicate posts that are joyful, relieved, or excited—for instance, people declaring that they have finally obtained the right treatment or achieved a sobriety milestone. These

communities' emphasis on recurring difficulties rather than steady well-being is consistent with the bottom-right quadrant (high valence, low arousal), which is also comparatively under-populated and linked to peaceful contentment.

Systematic differences can be seen by projecting subreddit-level averages onto the Circumplex. Deep in the negative valence region and frequently with moderate arousal, communities centered around acute crises or severe depression reflect persistent distress and low energy. The unsettled nature of anxious moods is reflected in the somewhat increased arousal found in anxiety-oriented subreddits at similar valence levels. Communities that focus on support and treatment had slightly lower levels of negativity and occasionally even modest positivity, which suggests that posts in these areas, which frequently deal with challenging subjects, contain more expressions of thankfulness, hope, and encouragement. Addiction treatment spaces are diverse; posts that celebrate progress and community support go toward higher valence, whereas posts regarding relapse tend to cluster in the negative valence region.

These patterns show that a basic four-emotion mapping can produce a complex affective landscape that distinguishes discourse types and communities. Additionally, they offer a condensed representation that can be connected to temporal trajectories or network measurements in subsequent research.

4.3 Subreddit Co-occurrence Network

The way people navigate Reddit's mental health ecosystem is summarized by the subreddit co-occurrence network. The network has a closely connected core of subreddits after isolating isolated communities and concentrating on the largest connected component. While edge widths show how many users appear in each connected communities, node sizes in the visualization correspond to degree centrality, or the number of other subreddits a community shares users with. Broad groups with titles indicating general mental health discussions, depression, anxiety, and treatment are examples of large, core nodes. These hubs provide users with access to a wide range of more specialized areas, such as those that are devoted to certain coping issues (e.g., addiction, urges to harm oneself), demographic groupings (e.g., women with autism), or diagnoses (e.g., OCD, ADHD, PTSD). The fact that many users interact with many communities in search of knowledge, validation, and specialized support is reflected in the complex network of edges surrounding this core.

The layout shows clusters. Mood and anxiety disorders that frequently co-occur clinically are included in one cluster; trauma-related

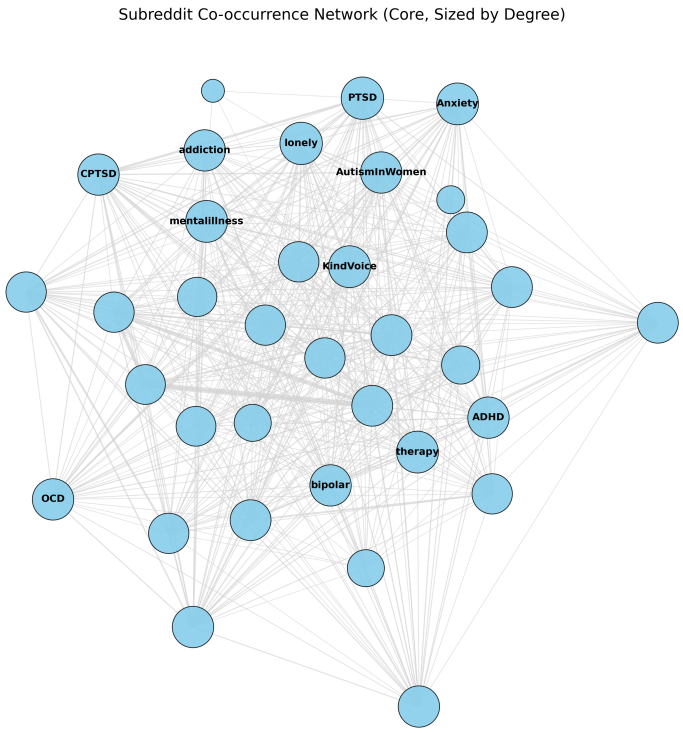


Figure 5: Subreddit Co-occurrence Network

and dissociation subreddits are included in another; addiction-related communities and sobriety support groups are included in a third cluster. Some users migrate between diagnostic areas and more general "venting" or advising forums, as indicated by the bridges connecting these to general support spaces. Perhaps because they serve highly narrow demographics, outlying nodes with fewer connections reflect niche subreddits that have little in common with one another.

According to an interpretation of this network from the perspective of mental health, Reddit serves as both a collection of specialized areas and a broader ecosystem where users move between communities as their needs and self-perceptions change. As they compile experiences from all around the network, highly linked hubs may be especially crucial for information flow and for identifying new distress patterns.

4.4 Emotion Transition Network

Discussions' emotional tone shifts are summarized by the thread-level emotion transition network. The resulting directed graph is straightforward enough to immediately comprehend because there are only four emotion categories. As anticipated, self-loops—melancholy followed by sadness and, to a lesser degree, optimism followed by optimism—are the most common transitions. These are collections of posts that share a same emotional tone across participants, such as a thread with several users discussing similar distressing experiences or a succession of supportive comments that build upon one another.

Transitions between emotions are more intriguing from the standpoint of support. There are many distinct edges, ranging from melancholy to hope or happiness. These usually relate to discussions in which an original poster conveys profound pain and replies provide understanding, consolation, or useful advice, changing the tone of the discussion to one of gratitude or hope. The positive change is further reinforced when the original poster responds with thanks in some instances. On the other hand, although they are less often, shifts from optimism or joy to despair do happen. These could represent circumstances in which a user's optimism is weakened by the unfavorable experiences of others, in which an initially positive update is greeted with skepticism, or in which additional discussion reveals unresolved distress. Anger-related transitions frequently occur in discussions of external institutions like healthcare, the workplace, or family disputes, when responses either reinforce or reframe frustration.

All things considered, the transition network emphasizes the persistence of negative affect in these societies as well as the significant con-

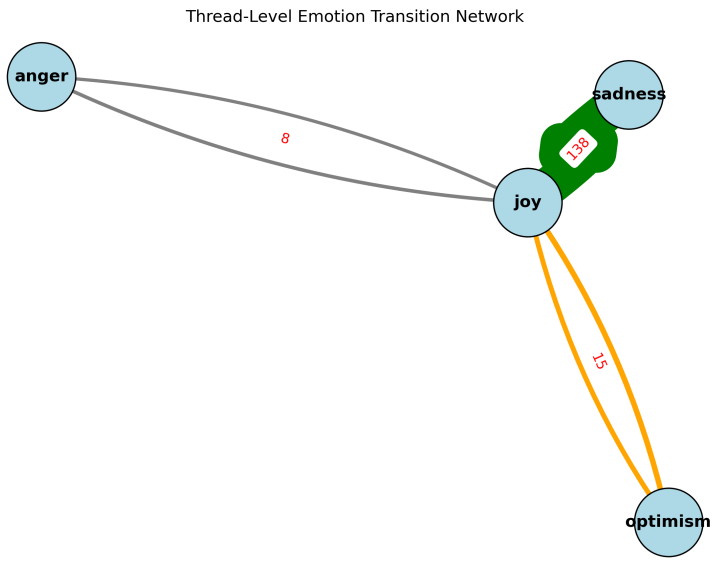


Figure 6: Emotion Transition Network

tribution that supportive relationships make to the introduction of more positive emotions. Despite the fact that the current analysis only approximates thread order and crude emotion categories, it shows that it is possible to predict emotional fluctuations in conversational structure as opposed to just individual postings.

4.5 Dashboard Illustration

The several analytical components are combined in an interactive manner in the Streamlit dashboard. The metrics panel usually displays a high amount of posts with moderate-to-high arousal and strongly negative average valence when a user chooses a subset of subreddits, including addiction and depression. The Circumplex scatter shows a dense cloud in the negative-valence, mid-arousal region, and the dominant emotion bar chart becomes strongly skewed towards melancholy. The dominance of melancholy scores for these communities is confirmed by KDE graphs.

The picture is altered by changing the filter to add more encouraging areas, like treatment or subreddits for general support. The Circumplex scatter broadens toward the right side of the plane, the percentage of optimism rises, and the average valence becomes less negative. Examples of posts chosen from these perspectives show various linguistic trends, such as the frequent use of positive language and the clear recognition of advancement.

Thus, the dashboard serves as a communication tool that enables non-

technical stakeholders to comprehend complex model outputs through visual exploration, as well as a validation tool that permits manual analysis of postings underlying certain patterns.

5 Discussion and Limitations

When combined, the findings show how useful it is to use transformer-based emotion modeling, continuous affective mapping, and network analysis to examine Reddit talk pertaining to mental health. When emotions are classified using a general RoBERTa model, the results show distinctions between communities that closely match to their declared focus and are in good agreement with intuitive assessments of the content. By converting these discrete probabilities into a two-dimensional affective representation that encompasses both polarity and activation, the Circumplex mapping reveals that a significant portion of the discourse in these subreddits is marked by a range of arousal and significantly negative valence.

The subreddit co-occurrence network adds a relational dimension, demonstrating that users frequently move between different spaces rather than living in isolated communities. This migration could be the result of co-morbid diagnoses, evolving self-perceptions, or just investigating alternative support options. In this network, hubs like general mental health and depression forums are crucial because they link more specialized communities and may act as early points of contact where emerging discomfort can be identified.

The emotion transition network provides an initial glimpse of how emotions change within threads at the conversational level. The network implies that negative emotions, especially melancholy, are sticky but not unchangeable, despite being constrained by the coarse four-emotion framework. As users receive encouraging comments, many threads show shifts from melancholy to joy or optimism, which is in keeping with earlier qualitative studies on online peer support. On the other hand, some transitions intensify negative emotions, emphasizing the possibility that uncontrolled environments may potentially heighten anxiety or pessimism.

These levels are connected via the interactive dashboard, which also shows how these analyses can be turned into action. It offers a practical means for researchers to investigate affective maps and emotion distributions within populations, identify anomalies, and formulate theories. A more developed version of this tool might help moderators or mental health professionals keep an eye on emotional climates or spot communities where discomfort is getting worse.

Crucially, this effort does not go so far as to develop an automated risk classifier or diagnostic system. Rather, it places a strong emphasis on framework development and descriptive analytics. However, if ethical and privacy concerns are properly taken care of, the same pipeline—data collecting, cleansing, transformer-based emotion modeling, affective mapping, and network construction—could serve as the foundation for early-warning technologies that highlight anomalous patterns for human assessment.

A number of constraints need to be noted. First, the dataset is confined by Reddit's API and the chosen sampling strategy: only up to 1,000 posts per subreddit and only top-level comments were collected. In addition to ignoring deeper reply chains that might exhibit distinct interaction and emotional dynamics, this underrepresents highly active communities. The emphasis on posts published after 2023 enhances recentness but leaves out longer-term changes in standards or responses to significant outside events. Second, domain mismatch is introduced when a generic RoBERTa model trained on Twitter is used. Sarcasm, mixed emotions, and precise clinical wording may be misclassified since it is not optimized for mental-health discourse or Reddit-specific language, although capturing broad emotional structure. Third, the mapping from four discrete emotions to valence–arousal coordinates is heuristic and coarse relative to the richness of true affective experience. Lastly, while the project as a whole raises ethical concerns about privacy, content sensitivity, and the dangers of overinterpreting automated inferences about mental health, the network analyses rely on simplifying assumptions (e.g., treating all shared users equally and approximating thread order).

6 Conclusion

This project created and put into use a multi-layered framework for analyzing Reddit discourse pertaining to mental health. The pipeline began by gathering raw posts and comments from a selected group of mental-health and support subreddits using PRAW. It then conducted a thorough cleaning and exploratory analysis before using an emotion classifier based on RoBERTa to assign probabilistic emotion profiles to each post. Each post was given a continuous emotional representation after these profiles were mapped into the Circumplex Model of Affect's valence–arousal space. Following network analyses that modeled emotional shifts within threads and connected communities via common users, the results were integrated into an explorable interface via an interactive Streamlit dashboard.

While the findings show pockets of optimism and joy linked to support and recovery narratives, they also underscore the prevalence of melancholy and negative valence in these networks. Both low-energy hopelessness and high-energy panic or rage are reflected in posts that cluster in negative regions with varied arousal, according to the valence–arousal mapping. The emotion transition network indicates that although negative affect is enduring, conversations can and frequently do change towards more positive feelings through encouraging comments, while the subreddit co-occurrence network reveals a closely knit core with powerful hubs.

6.1 Future Work

By implementing domain-specific transformers, adding symptom-level detection, and monitoring individual emotional trajectories throughout the network, future research might improve this approach. From a practical standpoint, these insights may allow clinician-guided, ethical dashboards to serve as early warning systems for increasing community distress. In the end, this experiment shows that integrating network analysis, continuous emotional modeling, and transformer-based categorization provides a workable method for comprehending the dynamics of online mental health. It provides a solid basis for examining how emotional climates change in public digital places by demonstrating that a rich, multi-scale representation of digital support and distress is achievable.

Link to the code:- https://github.com/VaibhavRawat39/

7 Acknowledgement

I would like to express my profound appreciation to Dr. Indu Bala for her outstanding leadership throughout this research endeavor. The steady hand that guided this effort to completion was her mentorship.

In particular, I would want to express my gratitude for the tremendous benefit of our weekly meetings. These meetings were essential to keeping the research moving forward and my progress on course. I am incredibly appreciative of her patience in answering my questions; she was always willing to clear up my confusion and give me the theoretical and technical clarification I required. Her commitment as a true educator was evident in her desire to go over ideas with me several times until I fully understood them.

I appreciate all of your time, patience, and priceless insights, Dr. Bala. Without your unwavering encouragement and support, our endeavor would not have been possible.

References

1. Russell, J.A. (1980) 'A circumplex model of affect', *Journal of Personality and Social Psychology*, 39(6), pp. 1161–1178. Available at: https://psycnet.apa.org/record/1981-25012-001

2. Bradley, M.M. and Lang, P.J. (1999) *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida. Available at: https://csea.phhp.ufl.edu/media/anewmessage.html

3. Russell, J.A. and Barrett, L.F. (1999) 'Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant', *Journal of Personality and Social Psychology*, 76(5), pp. 805–819. Available at: https://psycnet.apa.org/record/1999-10161-005

4. Yates, A., Cohan, A. and Goharian, N. (2017) 'Depression and self-harm risk assessment in online forums', in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pp. 1–12. Available at: https://aclanthology.org/W17-3101.pdf

5. Shing, H.C. et al. (2018) 'Expert, crowdsourced, and machine assessment of suicide risk via online postings', in *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pp. 25–36. Available at: https://aclanthology.org/W18-0603.pdf

6. Chancellor, S. and De Choudhury, M. (2020) 'Methods in predictive techniques for mental health status on social media: a critical review', *npj Digital Medicine*, 3, 43. Available at: https://www.nature.com/articles/s41746-020-0233-7

7. Schønning, V. et al. (2020) 'Social media use and mental health and well-being among adolescents: a scoping review', *BMJ Open*, 10(9), e031105. Available at: https://bmjopen.bmj.com/content/10/9/e031105

8. Wongkoblap, A., Vadillo, M.A. and Curcin, V. (2017) 'Researching mental health disorders in the era of social media: systematic review', *Journal of Medical Internet Research*, 19(6), e228. Available at: https://www.jmir.org/2017/6/e228/

9. De Choudhury, M., Gamon, M., Counts, S. and Horvitz, E. (2013) 'Predicting depression via social media', in *Proceedings of the 7th*

International AAAI Conference on Weblogs and Social Media (ICWSM), pp. 128–137. Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14432

10. Wang, Y. et al. (2023) 'Multi-modal deep-attention-BiLSTM based early detection of mental health issues on social media', *Scientific Reports*, 13, 22030. Available at: https://www.nature.com/articles/s41598-025-19141-0

11. Devlin, J. et al. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of NAACL-HLT 2019*, pp. 4171–4186. Available at: https://arxiv.org/abs/1810.04805

12. Liu, Y. et al. (2019) 'RoBERTa: A robustly optimized BERT pretraining approach', *arXiv preprint*, arXiv:1907.11692. Available at: https://arxiv.org/abs/1907.11692

13. Wolf, T. et al. (2020) 'Transformers: State-of-the-art natural language processing', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Available at: https://arxiv.org/abs/1910.03771

14. Alsentzer, E. et al. (2019) 'Publicly available clinical BERT embeddings', in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78. Available at: https://aclanthology.org/W19-1909.pdf

15. Ji, S. et al. (2022) 'MentalBERT: Publicly available pretrained language models for mental health text mining', *arXiv preprint*, arXiv:2205.03670. Available at: https://arxiv.org/abs/2205.03670

16. Barbieri, F. et al. (2020) 'TweetEval: Unified benchmark and comparative evaluation for tweet classification', in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650. Available at: https://aclanthology.org/2020.findings-emnlp.148.pdf

17. CardiffNLP (n.d.) *twitter-roberta-base-emotion*. Hugging Face model card. Available at: https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion

18. Demszky, D. et al. (2020) 'GoEmotions: A dataset of fine-grained emotions', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054. Available at: https://aclanthology.org/2020.acl-main.372.pdf

19. Mohammad, S.M. and Turney, P.D. (2013) 'Crowdsourcing a word–emotion association lexicon', *Computational Intelligence*, 29(3), pp.

436–465. Available at: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

20. Boe, N. et al. (2024) *PRAW: The Python Reddit API Wrapper – Official Documentation, Version 7.x.* Available at: https://praw.readthedocs.io/

21. Streamlit, Inc. (2024) *Streamlit Documentation, Version 1.x.* Available at: https://docs.streamlit.io/

22. Hagberg, A., Schult, D. and Swart, P. (2008) 'Exploring network structure, dynamics, and function using NetworkX', in *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, pp. 11–15. Available at: https://conference.scipy.org/proceedings/scipy2008/paper$_2$/$full_text.pdf$