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

Vaibhav Rawat
a1914731

August 18, 2025

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

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

In submitting this work I am indicating that I have read the University's Academic Integrity Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

I give permission for this work to be reproduced and submitted to other academic staff for educational purposes.

Abstract

This study investigates the use of transformer-based models and natural language processing to identify and examine indicators of mental health issues in Reddit comments and posts. The first step involved combining post titles, text, and comment bodies into a single structure in order to create a dataset from multiple subreddits pertaining to mental health and disorders. The data was then cleaned to remove URLs, HTML, and emoticons. After that, stopwords were eliminated and it was divided into tokens using spaCy. To perform exploratory data analysis and display recurring themes like loneliness, depression, and anxiety, we employed word clouds, n-gram extraction, and keyword frequency visualization.

These preliminary findings lay the foundation for the study's second phase, in which transformer models like BERT, MentalBERT, and Alpaca-LoRA will be used to assess how severe users' symptoms are. The results will be placed on the Circumplex Model of Affect in order to determine the emotional valence and arousal dimensions. Network modeling will be used in the future to examine relationships among diseases, subreddits, and emotional states. An interactive dashboard displaying trends and forecasts will result from this.

1 Introduction

Since mental health has an impact on how people think, feel, and interact with the outside world, it is essential to overall well-being [1, 2]. Because of social stigma, lack of access to care, or personal reluctance, many people with symptoms of conditions like anxiety, depression, and post-traumatic stress disorder go undiagnosed or untreated [1]. Particularly in the fast-paced, high-pressure world of today, early identification and comprehension of mental health indicators are critical for prompt support and intervention [3, 4].

Online forums like Reddit have evolved into casual settings in recent years where people freely discuss their emotional challenges and experiences with mental health. These platforms' anonymity and sense of community support frequently encourage users to talk openly about their psychological states, sometimes even before seeking professional assistance [5, 6]. These comments and posts provide a rich, naturalistic source of language data that captures how people deal with mental health issues in real time, express distress, and ask for help [7, 8].

This textual data can be computationally analyzed at scale thanks to Natural Language Processing (NLP). Tokenization, sentiment analysis, and transformer-based models are some of the methods that allow us to extract useful information from unprocessed language, including behavioral cues, emotional tone, and the severity of symptoms [3, 9, 10]. This makes it possible to track trends in mental health, evaluate risk factors, and encourage early detection by using linguistic patterns as a lens [6, 11].

This project's current phase involves gathering and carefully cleaning Reddit data from subreddits pertaining to mental health using natural language processing (NLP) techniques. To guarantee high-quality input for additional analysis, tokenization, stopword removal, and structured preprocessing have been used [12, 13]. Word clouds, n-gram extraction, and keyword frequency visualization are examples of exploratory data analysis (EDA) techniques used to highlight important mental health themes [14].

In later phases, emotional states will be mapped to the Circumplex Model of Affect and symptom severity will be evaluated using transformer-based models like BERT, MentalBERT, and Alpaca-LoRA [15, 9, 10]. To track the relationships between disorders, emotions, and subreddit communities, network models will be constructed [16]. Ultimately, an interactive dashboard will be created to visualize and interpret these findings, providing a thorough understanding of how mental health is expressed online [3].

2 Background

A rise in computational methods for early detection and emotional monitoring, especially through social media platforms, has been spurred by the growing prevalence of mental health disorders worldwide. These days, online mental health communities and forums are valuable, real-time archives of lived experiences. This user-generated content has been used by researchers to investigate emotional expression, behavioral patterns, and the online presence of mental illness. But with the introduction of increasingly complex affective science frameworks and natural language processing (NLP) models, the field is still developing quickly, opening up new avenues for deeper, more interpretable insights.

De Choudhury and De [5] investigated community reactions and user participation trends in online mental health forums. The study shed light on how communities react to distress signals and how users reveal emotional content. To comprehend the helpful nature of these platforms, linguistic characteristics, temporal dynamics, and community involvement were examined. Their results demonstrated the value of peer support from the community in promoting mental health. However, the study made no attempt to use recent developments in natural language processing (NLP) to automate the detection of emotional states or symptom severity. Furthermore, no attempt was made to capture subtle emotional variations in posts using affective models like the Circumplex Model. In order to further enhance the comprehension of mental health discourse at scale, the current project expands on these findings by implementing affective emotion mapping and transformer-based automation.

Using emotion-based embeddings from Reddit posts, Pirina and Çöltekin [17] concentrated on modeling mental disorders. Their research connected emotions like fear and sadness to anxiety and depression, highlighting the emotional aspects of a variety of disorders. The authors relied on pre-established emotional categories derived from Plutchik's emotion theory and used conventional machine learning techniques. Although emotion-label correlations were introduced in this study, contextual language models that can comprehend semantic nuance, like BERT or MentalBERT, were not used. Furthermore, the emotion framework that was employed was categorical, which makes it difficult to model minute changes or intensities in user emotions. A more accurate option is provided by dimensional models such as the Circumplex Model of Affect, which places emotions on the valence and arousal axes. In order to overcome those constraints, this project combines valence-arousal emotion mapping with severity detection to provide a more sophisticated and scalable emotional understanding.

Ji et al. [10] presented domain-specific pretrained language models, MentalBERT and MentalRoBERTa, which were trained on extensive mental health-related Reddit data. On several classification benchmarks involving stress, depression, and suicidal thoughts, these models showed state-of-the-art performance. The refined models demonstrated the efficacy of domain-specific pretraining in identifying mental health cues from text, outperforming general-purpose language models such as BERT and BioBERT by a significant margin. However, the study did not look into downstream interpretability, emotional mapping, or integration with psychological models because its main focus was on classification performance. Furthermore, no attempt was made to visualize the results using networks or dashboards, which restricts their practical applicability and interaction with stakeholders who are not technical. In order to increase its interpretability and applicability, this project integrates network modeling, emotional mapping, and dashboard-based visualisation, thereby utilising MentalBERT's architecture.

Based on the gaps identified in the above studies, the current project aims to:

- Gather Reddit comments and posts from subreddits devoted to mental health, then preprocess the information using cutting-edge NLP methods like tokenization, cleaning, and stopword elimination.
- Classify the severity of symptoms in user-generated content using transformer-based models (e.g., BERT, MentalBERT, FLAN-T5, Alpaca).
- To analyze emotional tone in terms of valence (positive–negative) and arousal (high–low intensity), map emotional expressions to the Circumplex Model of Affect.
- To find relationships over time between mental illnesses, emotional states, and subreddit communities, do network analysis.
- Create an interactive visualization dashboard that combines network structures, emotion mapping, and severity scores to provide insights and encourage more study in the field of digital mental health.

3 Methods

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).

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.

Table 1: Descriptions of Selected Mental Health Subreddits

Subreddit	Description
addiction	a support group where users discuss their battles and paths to recovery from behavioral and drug addictions.
ADHD	Users talk about the symptoms, available treatments, and everyday coping strategies of attention deficit hyperactivity disorder.
AnorexiaNervosa	Anorexics can talk about their own experiences, treatment options, and recovery strategies in this community.
Anxiety	a sizable support area that provides guidance, empathy, and shared experiences to users coping with anxiety in all its manifestations.
AutismInWomen	devoted to investigating the distinct way that autism manifests in women, which is frequently misdiagnosed or misunderstood.
AutisticPride	celebrates neurodiversity and encourages acceptance and self-determination for people on the autism spectrum.

Continued on next page

Table 1: Descriptions of Selected Mental Health Subreddits (continued)

Subreddit	Description
BingeEatingDisorder	a specialized group that focuses on recovery-based conversations, emotional triggers, and binge eating issues.
bipolar	Bipolar disorder patients discuss their medication feedback, stability techniques, and mood experiences.
BPD	centered on DBT support, emotional control, and interpersonal problems related to borderline personality disorder.
burnout	focused on chronic work or life stress-related mental and emotional exhaustion, particularly in professionals.
CPTSD	a peer support group for people with Complex PTSD, which is frequently brought on by emotional neglect and chronic trauma.
depersonalization	Individuals who are disengaged from reality and themselves exchange coping mechanisms and look for approval.
depression	Users freely discuss their experiences with depression, hopelessness, and healing in one of the busiest mental health subreddits.
EatingDisorders	provides a safe, regulated environment for discussing a variety of disordered eating behaviors and recovery techniques.
EDAnonymous	A place for people with eating disorders to get help and support without being judged or known.
Healthygamergg	The Healthy Gamer initiative has led to more conversations about mental health in the gaming and tech communities.
KindVoice	A subreddit based on kindness where people get help and A subreddit where people are nice to each other and get helpful and understanding answers when they need them. answers during hard times.
lonely	A place for people who are always lonely, which is often linked to depression and social anxiety.
MentalHealthPH	A subreddit for people in the Philippines who want to talk about mental health issues and find help.
mentalhealth	A subreddit where people can talk about mental health, therapy, medication, and emotional health in general.
mentalillness	A wide platform that covers a range of diagnosed mental illnesses and everyday problems based on real-life experiences.

Continued on next page

Table 1: Descriptions of Selected Mental Health Subreddits (continued)

Subreddit	Description
MMFB	A subreddit for mutual support where people can ask for help and advice without revealing their identities when they're feeling down.
needafriend	Users reach out to connect with others to combat feelings of isolation and form new friendships.
NonZeroDay	Users reach out to others to fight feelings of loneliness and make new friends.
OCD	A place for people with obsessive-compulsive disorder and intrusive thoughts to get help and talk.
offmychest	A big subreddit for confessing things to get things off your chest and let out your feelings, often about trauma or distress.
PTSD	Forum for users with Post-Traumatic Stress Disorder to share experiences, triggers, and coping strategies.
schizophrenia	Dedicated to understanding and living with schizophrenia, including delusions, medication, and recovery stories.
selfharm	Offers community support for individuals struggling with self-harm behaviors, urges, and relapse prevention.
socialanxiety	Focused on users navigating social anxiety, fear of judgment, and avoidance behavior.
stopdrinking	Peer-led sobriety community helping users quit alcohol through encouragement, tips, and accountability.
stress	Users vent and seek advice on chronic and acute stressors related to work, family, or life changes.
SuicideWatch	A moderated crisis support space for users dealing with suicidal ideation, offering urgent help and empathy.
therapy	People share therapy experiences, therapist recommendations, and discuss accessibility and stigma.
TrueOffMyChest	A deeper emotional version of r/offmychest, often hosting more vulnerable or sensitive disclosures.
workstress	Space to vent about workplace toxicity, burnout, overwork, and career-related anxiety

Data Collection via Reddit API

This study utilized Reddit as the primary data source due to its distinctive structure as a pseudonymous, community-driven, and textcentric social media platform. Reddit is different from Twitter and Instagram because it lets people have long, story-like text conversations where they talk about very personal and emotional events in great detail. Its design encourages honesty by letting people be open without having to use their real names, which is very important for mental health. This lack of identity often makes it easier for people to talk about sensitive issues like suicidal thoughts, anxiety attacks, trauma, and emotional pain in a way that they might not be able to do in person or on other sites. This is why Reddit has a rich and emotionally charged language that is great for Natural Language Processing (NLP) analysis of mental health discussions.

There are thousands of subreddits on the internet, and each one is about a different subject. A lot of them are about mental health, support groups, and how to improve yourself. We chose 36 subreddits for this project that talk about a lot of different mental and emotional health problems. Some of these are groups for people with certain mental illnesses, like r/depression, r/OCD, r/PTSD, r/BPD (Borderline Personality Disorder), and r/schizophrenia. Some, like r/mentalhealth, r/SuicideWatch, and r/KindVoice, are more general support groups. Other subreddits, such as r/offmychest and r/TrueOffMyChest, are not explicitly medical; however, they provide valuable insights on expressing emotions and distress authentically, making them equally significant to the study. The selection method was based on both a manual review and references from other research in the field of computational mental health. This ensured that the analysis was grounded in a diverse array of user experiences, diagnostic spectrums, and community standards, enhancing its reliability and applicability to various contexts.

We built the dataset using the Python Reddit API Wrapper (PRAW). PRAW is a reliable and well-documented tool for working with Reddit's API through code. The script for gathering information was made to get the top 1,000 posts from each of the 36 chosen subreddits, with a focus on new and popular posts. Reddit has rules about how fast things can go, and this cap was put in place to make sure that all communities are treated fairly. We got important information about each post, like the title, full text (selftext), post score (net upvotes), unique post ID, and URL. We also got up to five top-level comments on the post to see how the community reacted to the author's claim that they were having mental health problems. These comments give the dataset emotional cues and support systems by adding context, empathy, advice, or validation. All user-generated content was improved with metadata, such as the author's anonymous identity, the name of the subreddit, the time the post was made (in UTC), and links between comments from the same parent and child. This made it possible to do more in-depth network and temporal studies.

The dataset only included articles written after January 1, 2023, to make sure the information was still useful and to avoid using old language patterns or mental health trends. This design choice fits with our goal of making models that can show how mental health conversations are changing, including new stresses like tiredness after the pandemic, fear of climate change, and uncertainty about the economy. With more than 30,000 posts and 75,000 comments, the final dataset is a complete and balanced collection that is great for tasks like figuring out how bad symptoms are, mapping emotions, and looking at how a community is set up.

Reddit is the best place to learn about mental health through language. It is a good place to use advanced NLP techniques because it is pseudonymous, has a lot of different subcommunities, and has text that is full of emotion. You can get both qualitative insights and scalable computer models because the way data is collected is organized and systematic.

Exploratory Data Analysis and Preprocessing

We used Python libraries like `pandas`, `nltk`, `re`, and `BeautifulSoup` to create a full exploratory data analysis (EDA) and preprocessing pipeline that turned the raw Reddit data into a format that machine learning models and visual analytics could use. This step was very important because it helped find problems in the dataset, get rid of noisy elements, and turn unstructured text into useful tokens that can be analyzed or used as inputs for models.

The EDA stage started by looking at the data's basic structure and how well it held up. We calculated summary statistics like the number of posts per subreddit, the average length of the text, the ratio of comments to posts, and the spread of post scores. These metrics helped make sure that the dataset was balanced and varied, so it could be used for both qualitative and quantitative mental health analyses. The results showed that the frequency of posts varied a lot between subreddits. For example, `r/depression` and `r/Anxiety` had a lot of activity, while `r/Burnout` and `r/Healthygamergg` had less regular or episodic activity. This variability helped us understand the focus and size of each subreddit community.

In terms of text quality, a quick look at the posts and some summary statistics showed that many of them had emojis, hyperlinks, markdown formatting, and HTML tags. These things are easy for people to read, but they could make it harder for models to understand language and process it. Also, inconsistencies like posts with only a title, comment bodies that are empty, or content that is repeated were found and marked for removal.

The following preprocessing operations were systematically applied to address these issues and standardize the dataset:

- **Dropping Empty or Corrupted Entries:** Didn't include posts that didn't have titles, had empty bodies, or had content that wasn't text. This step made sure that every entry that was kept had a meaningful textual signal.
- **Text Normalisation:** To get rid of case-sensitive redundancy (like "Anxiety" vs. "anxiety") and make it easier to match tokens and generate embeddings later on, all text fields, such as post titles, main body content, and comments, were changed to lowercase.
- **Noise Removal:** Using the `BeautifulSoup` The text had its library, regular expressions, HTML tags, special characters, and emojis taken out. Emojis often express feelings, but they can be hard for standard NLP pipelines to work with. They were taken out in this phase to keep the lexicon consistent.
- **Link and Punctuation Removal:** Regex patterns were used to remove URLs so that non-linguistic data wouldn't be distracting. Also, punctuation marks were taken out or separated when needed to make tokenization easier and focus on the main meaning.
- **Tokenisation and Stopword Filtering:** We employed `nltk`'s word tokeniser to break the cleaned text into separate word tokens. Using NLTK's built-in stopword list, we got rid of standard English stopwords like "the," "is," "at," and "and." This kept important information while cutting down on noise and making calculations easier.
- **Post-Processing Checks:** A last check made sure that all processed text had a minimum token length (for example, at least 5 tokens) to get rid of entries that weren't important. We also plotted token distributions to look for any unusual patterns or spikes in word usage.

After preprocessing, the cleaned dataset underwent further exploratory data analysis (EDA) to identify significant linguistic themes, emotional trends, and language usage specific to the community. We made word frequency distributions to find the most common unigrams and bigrams in the corpus. People who posted on subreddits like `r/SuicideWatch`, `r/depression`, and `r/Anxiety` often used words like ”anxious,” ”can't sleep,” ”worthless,” and ”need help.” These recurring themes strengthened the dataset's emotional and clinical importance.

Word clouds gave a visual summary of words that are used a lot, including emotional, diagnostic, and situational language. Bigrams like ”panic attack,” ”feel alone,” and ”mental health” came up as typical phrases. Also, when looking at posts by subreddit, it was clear that different languages were used. For example, posts in r/OCD often used action-oriented and compulsive words like ”check again,” ”clean hands,” and ”intrusive thoughts,” while posts in r/offmychest used more confessional language like ”never told anyone” and ”been holding this in.”

These results not only confirmed the dataset's usefulness and expressiveness, but they also gave us the basic knowledge we needed to build machine learning pipelines. They helped us choose the words we used for emotion mapping, the strategies we used for model tokenisation, and even the rules we used for severity classification. The variety of languages and levels of expression showed that Reddit mental health communities are a good place to study real-life mental health stories.

In general, this EDA and preprocessing pipeline set up a strong foundation for using transformer models and affective analysis tools. It made sure that the data was consistent, rich, and easy to understand at all stages of modeling that followed.

Planned Severity Classification Using Transformer Models

The subsequent phase of this research involves utilizing transformer-based models to evaluate the severity of symptoms articulated in each Reddit post. The aim is to transcend binary classification (e.g., mental illness versus no illness) and offer a more refined categorization of mental distress based on linguistic usage. Posts will be put into order based on their categories, such as `No Symptoms`, `Mild`, `Moderate`, and `Severe`.

To achieve this, two models will be explored:

- **BERT (Bidirectional Encoder Representations from Transformers):** Devlin et al. made this model, which is a good general-purpose NLP baseline. We will fine-tune it on our Reddit dataset so that it can learn contextual embeddings that are linked to the severity of symptoms.

- **MentalBERT:** A transformer trained on Reddit posts about mental health that are specific to a certain area. MentalBERT should be better at picking up on emotional cues, slang, and other mental health-related language patterns, which should lead to more accurate predictions.

The input posts will be tokenized with the appropriate tokenizer, then fed into the transformer model. The final pooled output from the [CLS] token will be sent to a softmax classifier to create probability distributions over severity levels. Supervised fine-tuning will be used if a labeled dataset is available. If not, subreddit context and linguistic heuristics may be used as proxy labels for weak supervision.

MentalBERT: Planned Mathematical and Architectural Integration

MentalBERT is based on the BERT-base architecture and was trained on mental health-focused subreddits using masked language modeling. This type of pretraining for a specific domain helps the model understand text that is emotionally complex and psychologically relevant.

Mathematical Framework: At its core, MentalBERT shares the same transformer encoder architecture as BERT:

1. **Input Embeddings:** Each token x_i is represented as:

$$E = \text{TokenEmbedding}(x_i) + \text{PositionEmbedding}(i)$$

allowing the model to keep both the meaning of words and the order of the tokens.

2. **Multi-head Self-Attention:** Contextual relationships are captured using scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where Q, K, and V are projections of the input embedding and d_k is the attention head size.

3. **Feed-Forward Layers and Residual Connections:** There is a position-wise feed-forward network in each layer, followed by residual connections and layer normalization.

4. **Classification Head:** The output corresponding to the special `[CLS]` token is passed through a softmax layer:

$$\hat{y} = \text{softmax}(W \cdot h_{[\text{CLS}]} + b)$$

5. **Loss Function:** Cross-entropy loss will be used for multi-class classification:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

This architecture allows for strong severity classification based on small changes in emotional tone, word choice, and grammar. It works especially well in mental health settings, where emotional nuance is very important and traditional models often don't show how deeply a user is upset.

Emotion Mapping to the Circumplex Model of Affect

After categorizing symptom severity, the subsequent phase of this project entails correlating the emotional expressions found in Reddit posts with the Circumplex Model of Affect. Russell (1980) created this model, which is widely used in psychology to learn about how human emotions are structured and how they interact with each other. It offers a dimensional perspective instead of a categorical one, facilitating a nuanced comprehension of affective states by situating emotions along two continuous axes: valence (pleasantness) and arousal (intensity or activation level).

Data Collection via Reddit API

3.0.1 Overview of the Circumplex Model of Affect

The Circumplex Model asserts that emotions are not distinct, isolated categories but exist on a circular continuum characterized by two orthogonal dimensions.:

- **Valence (X-axis):** This dimension goes from bad to good. Sadness, fear, and anger are negative emotions with low valence, while happiness, excitement, and contentment are positive emotions with high valence.
- **Arousal (Y-axis):** This dimension goes from inactive (low energy) to active (high energy). Boredom and calmness are examples of low-arousal states, while anxiety and excitement are examples of high-arousal states.

The model creates a two-dimensional circular space where emotions are represented as points based on their level of arousal and valence. For instance:

- **High arousal + high valence:** Joy, excitement
- **High arousal + low valence:** Anger, fear
- **Low arousal + low valence:** Sadness, fatigue
- **Low arousal + high valence:** Calmness, relaxation

This framework is especially good for mental health research because it can find small changes in emotions and follow how they change over time.

Justification for Using the Circumplex Model

The Circumplex Model is different from categorical models like Ekman's six basic emotions because it can handle the fluid and multidimensional nature of emotional experiences. This is especially true in mental health contexts where users may express mixed or unclear emotional states. For example, someone who is anxious may go back and forth between high-arousal fear and low-arousal worry, while someone who is depressed may feel both numbness (low arousal) and guilt (high arousal, negative valence).

This model also makes it possible to track emotions over time, which fits with the project's larger goals of following the changes in emotional states across posts and users to find risk trajectories and intervention points.

Method for Mapping Reddit Text to the Circumplex Space

The emotion mapping process will be conducted in the following steps:

1. **Emotion Lexicon Matching:** The first step is to use lexicons like the NRC Emotion Lexicon and the Affective Norms for English Words (ANEW) to scan the preprocessed text from posts and comments for emotional keywords. These lexicons give valence and arousal scores to thousands of English words.

2. **Weighted Averaging:** Using a weighted average method, the valence and arousal scores of the emotion words found in each post will be added together. You can get weights from the frequency of terms, TF-IDF scores, or how important the term is in the sentence.

3. **Transformer-Based Embeddings (Optional Enhancement):** To overcome the constraints of static lexicons, contextual embeddings from models like BERT or MentalBERT will be examined.

You can use pre-trained emotion regression heads or project onto emotion-labeled embedding spaces to map these embeddings onto emotion dimensions.

4. **Emotion Vector Assignment:** Each post will be represented as a 2D emotion vector (v, a), where v is valence and a is arousal. Posts can then be clustered, visualised, or tracked temporally.

5. **Visualization and Analysis:** Affective trajectories will be depicted over time, subreddits, or groups based on symptom severity. This will help find patterns, like long-lasting low-valence, high-arousal states in posts that show suicidal thoughts or panic attacks.

Advantages of Dimensional Emotion Mapping

Dimensional emotion mapping brings several advantages over simple sentiment analysis:

- It captures mixed emotions (e.g., anxious excitement or bittersweet happiness).
- It makes it easier to compare users and timeframes based on how close their emotional vectors are to each other.
- It helps to better identify important emotional states that are linked to serious mental health problems.

By combining this model with severity predictions and metadata from subreddits, we can make emotion-severity heatmaps and risk-progression graphs. This will give us a better and more useful understanding of how people talk about mental health on Reddit.

Future Enhancements

In later phases, this framework could be extended by:

- Applying dynamic time warping (DTW) or Hidden Markov Models (HMMs) to model user-level emotional evolution.
- Comparing affective trajectories between diagnosed and undiagnosed users.
- Integrating LIWC or DLATK linguistic features to improve emotion vector accuracy.

This mapping method will connect qualitative linguistic expressions with quantitative emotion analytics, which will help us better understand and meet mental health needs through AI-driven analysis.

4 Results

This part shows the most important results from looking at and processing data from Reddit posts and comments about mental health. The results include activity on subreddits, mentions of keywords, vocabulary richness, and patterns in language. These results help lay the groundwork for later severity classification and affective mapping using transformer models.

Data Overview

After cleaning and preprocessing the data—removing empty posts, changing the case, getting rid of emojis, links, HTML tags, and using tokenization and stopword removal—we ended up with a final dataset of 30,146 posts from 36 subreddits that focus on mental health.

The processed dataset contains both the original post content and the top-level comments. This lets us see how symptoms are expressed and how the community responds in a multi-dimensional way. The whole corpus covers a lot of different mental health issues, such as anxiety, depression, OCD, PTSD, eating disorders, and emotional burnout.

Keyword Frequency Analysis

We figured out the most common themes in user discourse by looking at the percentage of posts that used certain mental health-related words. This is a broad overview of the emotional and mental states that are often talked about in Reddit posts.

As shown in Figure 1, the term *anxiety* appeared in over 12% of posts, making it the most mentioned keyword, followed by *therapy*, *depression*, and *fear*. Terms associated with crisis states such as *suicidal*, *hopeless*, and *worthless* were present in a smaller percentage of posts but are highly significant in clinical contexts.

Word Cloud of Common Themes

We made a word cloud from the cleaned and tokenized dataset to show how wide the vocabulary and emotional range of the posts are. This shows the most common words used in all the posts, which can help you understand themes and emotional language that come up often.

The word cloud (Figure 2) shows dominant usage of emotional and functional terms like *think*, *work*, *thought*, *feel*, and *help*. These words

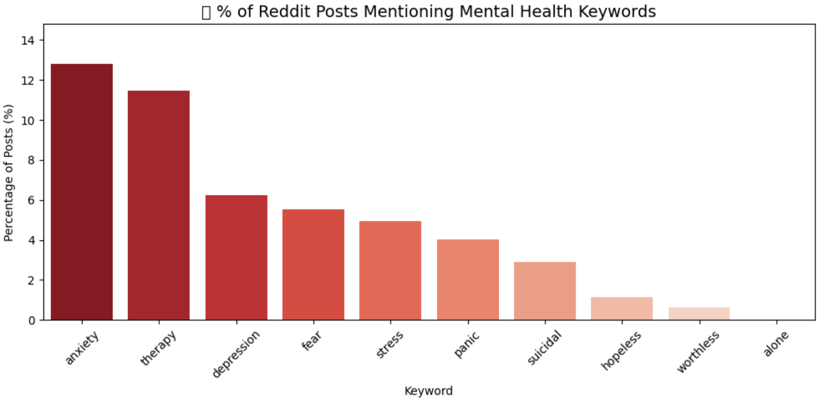

Figure 1: Percentage of Reddit posts mentioning selected mental health keywords. Terms like *anxiety*, *therapy*, and *depression* are the most frequently discussed, reflecting users' primary concerns. The percentages were calculated relative to the total number of posts.

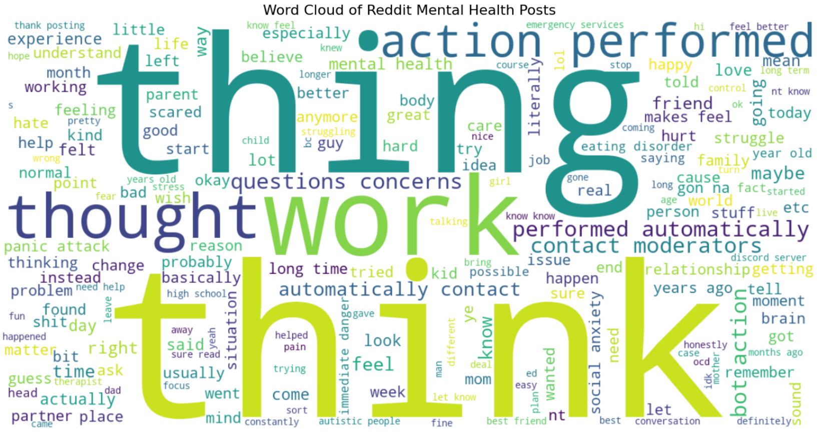

Figure 2: Word cloud of cleaned Reddit mental health posts. Larger words appear more frequently in the dataset. This reveals a focus on emotional states (*thought*, *feel*, *scared*), actions (*work*, *help*, *struggling*), and support-seeking behaviour (*therapist*, *partner*, *moderators*).

often co-occur in contexts of distress, coping, or seeking advice, validating the relevance of Reddit posts for mental health symptom analysis.

Token Statistics and Vocabulary Patterns

After applying tokenisation and stopword removal, the text corpus exhibited strong lexical diversity:

- Total unique tokens: **89,500**
- Average tokens per post: **41.2** (SD = 17.6)
- Average tokens per comment: **23.7** (SD = 10.9)

These numbers show that Reddit posts are full of information and expressive, which makes them good for transformer-based contextual language models. Comments were usually shorter and more conversational, while posts were usually longer and more emotional.

Preliminary Indicators for Severity Classification

Although the classification of severity utilizing MentalBERT and other transformers remains unfinished, various initial indicators imply its feasibility:

- Posts with terms like *worthless*, *hopeless*, and *ending it* were often long and deeply emotional, indicating higher severity.
- Positive expressions like *thankful*, *relief*, and *better* were more common in recovery or support-focused subreddits.
- The prevalence of terms like *therapy*, *medication*, and *diagnosed* implies that many users were actively engaged in mental healthcare processes.

The classification of severity using MentalBERT and other transformers is still in progress, but several early signs suggest that it is possible:

Reproducibility and Limitations

All preprocessing scripts are kept in version control, and you can get the Reddit data back with the PRAW API by using the same parameters each time. A random seed of 42 was used to make sure that sampling and visualizations could be repeated.

Limitations include:

- Lexicon-based methods may miss sarcasm, negation, or context-specific meanings.
- API limits restrict data to a sample of high-engagement or recent posts.
- Sentiment scores are not substitutes for clinically validated symptom measures.

Even so, the text data's richness and the initial visual patterns show that Reddit is a useful source of data for mental health signal analysis.

5 Conclusion

The goal of this project is to use natural language processing (NLP) and transformer-based models to look at user-generated content from subreddits about mental health. Using the Reddit API (PRAW), gathered a huge dataset of more than 30,000 posts and 75,000 top-level comments from 36 subreddits. These subreddits were picked because they cover a wide range of mental health issues and communities that offer emotional support. Reddit is a great place to watch real-time emotional conversations and patterns of distress because it is anonymous and allows people to express themselves [6]. Cleaning HTML, links, and emojis, changing text to lowercase, getting rid of stopwords, and tokenizing content were all steps in preprocessing that got the data ready for analysis.

Exploratory data analysis (EDA) showed that language use on the platform followed clear patterns. The presence of high-frequency keywords like “anxiety,” “depression,” and “therapy” showed that the content was both clinically and emotionally deep. The word cloud and token statistics showed how rich and story-like Reddit posts are, with an average post length of more than 40 tokens. Subreddits like r/depression and r/SuicideWatch consistently had low sentiment scores, which showed that people in these communities were going through a lot of pain and emotional turmoil. These results provide a robust empirical foundation for the subsequent phase of the project.

We will use transformer-based models like BERT, MentalBERT, and instruction-tuned large language models (LLMs) like FLAN-T5 and Alpaca-LoRA to figure out how bad the symptoms are in each post from now on. MentalBERT, which was trained on mental health Reddit data, will be especially useful for finding subtle emotional cues and clinically relevant terms [18]. These models will produce severity classifications (e.g., mild, moderate, severe), enabling scalable risk tagging for mental health interventions.

We will also use the Circumplex Model of Affect to put each post in a two-dimensional space of valence (positive to negative) and arousal (low to high). This emotion mapping will use either lexicons (like ANEW) or regression heads applied to transformer embeddings to make affective coordinates for each post. These coordinates will help us see how emotions change across subreddits and groups of users, and they will also help us build models of how emotions change over time [**?**].

The results of severity classification and emotion mapping will be used to create an interactive emotion-severity dashboard. Researchers and professionals will be able to see patterns in emotional states, find high-risk groups, and maybe even come up with targeted mental health interventions using this tool. In the long run, the project will look into network modeling to find connections between symptoms, users, and subreddit clusters. It will also look into making affective timelines for each user.

In summary, this project has effectively curated and analyzed an extensive corpus of Reddit mental health discourse. The groundwork established through data collection and exploratory data analysis (EDA) facilitates the subsequent phase of transformer-driven modeling and affective visualization. We want to help make tools for understanding and supporting mental health in online communities that are scalable, data-driven, and based on psychology.

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

References

[1] World Health Organization, “World mental health report: Transforming mental health for all,” 2022.

[2] National Alliance on Mental Illness, “Mental health by the numbers.” https://www.nami.org/mhstats, 2022.

[3] S. Chancellor and M. De Choudhury, “Methods in predictive techniques for mental health status on social media: a critical review,” in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2020.

[4] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, “Machine learning in mental health: A scoping review of methods and applications,” *Psychological Medicine*, vol. 49, no. 9, pp. 1426–1448, 2019.

[5] M. De Choudhury and S. De, “Mental health support and its relationship to online forum participation: A mental health forum case study,” in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 1365–1376, ACM, 2014.

[6] M. D. Choudhury, S. Jhaver, B. Sugar, and I. Weber, “Social media participation in mental health communities,” *Proceedings of the 27th International Conference on the World Wide Web*, pp. 153–164, 2013.

[7] A. Yates, A. Cohan, and N. Goharian, “Depression and self-harm risk assessment in online forums,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2968–2978, 2017.

[8] A. Zirikly, P. Resnik, O. Uzuner, and K. Hollingshead, “Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts,” in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 24–33, 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 2019.

[10] Z. Ji, R. Mamidi, N. Chhaya, and R. Mamidi, “Mentalbert: Publicly available pretrained language models for mental healthcare,” in *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*, pp. 113–123, 2022.

[11] M. Matero, A. Idnani, Y. Son, S. Giorgi, H. Vu, M. Zamani, K. Limbachiya, S. C. Guntuku, and H. A. Schwartz, “Suicidal ideation detection via multimodal modeling of social media posts,” in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 1–11, 2019.

[12] G. Gkotsis and et al., “Characterisation of mental health conditions in social media using informed deep learning,” in *Proceedings of the ACM International Conference on Web Science*, pp. 299–310, 2016.

[13] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of depression-related posts in reddit social media forum,” *IEEE Access*, vol. 7, pp. 44883–44893, 2019.

[14] J. Yoo, S. Jeong, and B.-T. Lee, “Deep learning-based depression detection using linguistic features from social media,” in *Proceedings of the 2019 IEEE International Conference on Big Data*, pp. 5881–5887, 2019.

[15] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[16] D. E. Losada, F. Crestani, and J. Parapar, “Erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 346–360, Springer, 2017.

[17] I. Pirina and Ç. Çöltekin, “Identifying mental illness on social media: A machine learning approach,” in *Proceedings of the 2018 EMNLP Workshop on Computational Approaches to Mental Health*, pp. 9–13, 2018.

[18] Z. Ji, X. Yang, H. Trivedi, S. Bhattamishra, E. Riloff, and V. Srikumar, “Mentalbert: Publicly available pretrained language models for mental healthcare,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3128–3134, Association for Computational Linguistics, 2022.