# LLMs for Mental Health Early Detection

a1907385 Chia-Hao Lo
2025/06/22

## 1. Introduction

Mental health issues are highly common worldwide. In Australia, nearly half of all adults (7.3 million people) are expected to experience a mental illness at some point. Early intervention often leads to better outcomes in treating mental health problems. However, early detection is still a major challenge because of many individuals delay seeking help until symptoms become severe.

Early signs of mental health struggles often show up as small changes in our daily routines, but they're easy to miss. Luckily, as we share snippets of our daily lives on social media, these platforms can reveal those subtle changes. These data contain a rich of information that allows us to monitor mental states. Some online platforms such as Reddit or Beyond Blue are the environments for people to let off steam. These forums offer a valuable, real-time window into emotional states. This makes them powerful resources for early mental health monitoring. Yet, analyzing such content is not straightforward, such as unstructured, diverse in tone and language, and often includes slang, abbreviations, or vague emotional expressions. Traditional analytical methods struggle to process and interpret these large volumes of informal data.

The appearance of large language models (LLMs), such as ChatGPT and domain-adapted models like BERT, has shown the powerful ability to analyze unstructured text by capturing context, sentiment, and linguistic nuance [5]. Yet, LLMs are inherently limited in reliability due to inconsistent accuracy and their black-box nature, which makes it difficult to explain or justify model outputs. Besides, LLMs are not well-suited for long-term user monitoring because they lack native support for time series modeling. As a result, while LLMs are powerful for extracting semantic features, they remain inadequate when used alone for early mental health prediction.

To solve these limitations, our approach combines the LLM model result with temporal processing. We begin with NLP preprocessing and feature extraction and followed by the use of MentalRoBERTa or FLAN-T5 to analyze post content. The outputs are mapped to the Circumplex Model of Affect to structure emotional states. Then we will feed into an LSTM model to observe long-term user patterns. Finally, we will apply LIME to provide interpretability for model predictions.

## 2. Background

### 2.1. Domain-Tuned LLMs and NLP for Mental Health Text Analysis

Studies have shown that the way people write can be strongly connected to their mental health [4]. Besides, unlike formal mental health assessments, online platform provides spontaneous, unfiltered, and time-series text. This brings researchers to observe how a individual expresses their emotions in the moment without any clinical intervention or diagnosis [3].

While general-purpose LLMs like ChatGPT are effective at understanding broad language patterns, they are not optimized for domain-specific tasks. This project instead uses MentalRoBERTa and FLAN-T5 because these two specialized models designed for more focused applications. MentalRoBERTa is based on BERT, a transformer model that learns contextual word representations by masking and predicting tokens. It has been fine-tuned on Reddit mental health data to make it more sensitive to psychological language. FLAN-T5 is part of Google's FLAN family of instruction tuned models, especially in zero-shot and few-shot tasks. It provides strong performance without requiring large labeled datasets, which makes it ideal for detecting emotional tone in resource constrained settings [2, 5].

To prepare text for these models, we apply standard Natural Language Processing (NLP) techniques. This includes text cleaning (removing emojis, links, and special characters), tokenization, lowercasing, stopword removal, and lemmatization. These steps reduce noise and standardize the data, ensuring that both LLMs and traditional feature extractors like LIWC and DLATK can process the input effectively.

### 2.2. Objective

- To extract and classify emotional states from mental health forum posts using domain-specific LLMs (MentalRoBERTa, FLAN-T5), mapped to the Circumplex Model of Affect.

- To model the temporal progression of depressive symptoms using LSTM-based architectures and identify early warning patterns. Including tracking comments.

- To apply interpretability techniques (e.g., LIME, attention mechanisms, GPT-generated summaries) to improve trust and explainability of model predictions.

### 2.3. Data Collecting & Preprocessing

This project will use one or both text data from two sources: Reddit and Beyond Blue. Reddit provides a large volume of informal, emotionally expressive posts from global users, it organized by mental health-related subreddits such as r/depression and r/anxiety. Its high user activity and diverse language make it suitable for training LLMs, though it requires careful filtering due to noise and lack of geographic specificity. In contrast, Beyond Blue offers a more structured and moderated dataset focused on Australian users. Posts are categorized by mental health topics and tend to be more supportive and direct, it makes them valuable for culturally grounded analysis.

For web scraping, we use a custom web scraping tool using Python, combining requests, BeautifulSoup, and Selenium to access both posts and replies. The scraper targets key categories including depression, anxiety, PTSD and trauma, and suicidal thoughts and self-harm. It retrieves posts made between May 2024 and May 2025 and in Australia region with associated metadata such as post time, author, comment count, and forum topic. The data will be saved in structured .csv files, separated into posts and comments for easier downstream processing.

## 3. Methods

This project implements a multi-stage approach combining traditional NLP, domain-adapted LLMs, temporal modeling, and interpretability tools to detect and track mental health signals in user-generated text.

### 3.1. NLP preprocessing &Feature Extraction

We begin by collecting forum and post data from Reddit and Beyond Blue. The raw text is processed using standard NLP techniques. Then we apply WordPiece tokenization to match our LLM vocabularies. For each post, we concatenate the LIWC vector and the DLATK category frequency vector with the token embeddings, then feed the combined representation into our LLM backbones.

### 3.2. LLM-Based Mental Health Signal Modeling

The cleaned text with the feature vector is then passed through MentalRoBERTa or FLAN-T5. For MentalRoBERTa, we will extract the [CLS] token embedded for the final layer as a dense post-level feature vector for LSTM input. [1]. We will use FLAN-T5 to generate structured outputs (e.g. risk levels or affective states), which are converted into numeric vectors for downstream modeling.

The LLM outputs a numeric feature vector for each post, capturing contextual, emotional, and psychological cues. These vectors such as classification logits, pooled hidden states, or affective embeddings—serve as inputs to the LSTM model described in the next section, which captures temporal patterns in users' mental health signals over time.

### 3.3. Temporal Modeling with LSTM

To observe how users' emotional states evolve over time, we build user-level timelines by chronologically organizing the post-level feature vectors extracted by the LLMs. These vectors - derived from both the fused text and psycholinguistic inputs - capture a range of emotional and psychological signals. We feed these sequences into an LSTM model, which learns temporal patterns and dependencies, enabling the detection of trends such as emotional deterioration or recovery.

### 3.4. Interpretability via LIME

To improve transparency and trust in predictions, we apply LIME (Local Interpretable Model-Agnostic Explanations). LIME emphasizes the most influential words or features in each classification. This can make sure researchers to understand and validate the model's reasoning behind mental health-related labels.

## 4. Conclusion

In summary, this project aims to develop a hybrid framework that combines domain-adapted LLMs, psycholinguistic analysis, and temporal modeling to detect and track mental health early signals in social media text. By leveraging Reddit and Beyond Blue data, we expect to generate interpretable, time-aware predictions of users' emotional states. The anticipated outcome is a system that not only identifies early signs of mental distress but also provides insight into emotional progression over time, supporting future efforts in early intervention and mental health monitoring.

## References


[1] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. MentalBERT: Publicly available pretrained language models for mental healthcare. In *International Conference on Language Resources and Evaluation*, page 7184–7190, 2022. 2

[2] Min Li, Dongxiao Gu, Rui Li, Yadi Gu, Hu Liu, Kaixiang Su, Xiaoyu Wang, and Gongrang Zhang. The impact of linguistic signals on cognitive change in support seekers in online mental health communities: Text analysis and empirical study. *Journal of Medical Internet Research*, pages 1–15, 2023. 1

[3] De Choudhury Munmun, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137, 2021. 1

[4] Andrea N Niles, Kate E Byrne Haltom, Catherine M Mulvenna, Matthew D Lieberman, and Annette L Stanton. Randomized controlled trial of expressive writing for psychological and physical health: the moderating role of emotional expressivity. anxiety stress coping. *Anxiety Stress Coping*, pages 1–17, 2014. 1

[5] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. Mental-LLM: leveraging large language models for mental health prediction via online text data. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, volume 8, pages 1–32, 2024. 1